MEASURE WORKSHEET

This document summarizes the evaluation of the measure as it progresses through NQF's Consensus Development Process (CDP). The information submitted by measure developers/stewards is included after the Brief Measure Information, Preliminary Analysis, and Pre-meeting Public and Member Comments sections.

To navigate the links in the worksheet: Ctrl + click link to go to the link; ALT + LEFT ARROW to return

Brief Measure Information

NQF #: 0033
Measure Title: Chlamydia Screening in Women (CHL)
Measure Steward: National Committee for Quality Assurance
Brief Description of Measure: The percentage of women 16–24 years of age who were identified as sexually active and who had at least one test for chlamydia during the measurement year.
Developer Rationale: This measure assesses the percentage women 16-24 years of age who were identified as sexually active and who received a test for chlamydia. The improvement in quality envisioned by the use of this measure is increased identification of untreated chlamydia infections in women that can lead to serious and irreversible complications, and can be unknowingly transmitted to sexual partners. Despite the availability of effective treatments, a large proportion of sexually active individuals continue to go undiagnosed due to the disease's asymptomatic nature. Early detection, screening, and treatment have proven to be effective in managing and preventing chlamydia.

Numerator Statement: Females who were tested for chlamydia during the measurement year.
Denominator Statement: Females 16-24 years who had a claim or encounter indicating sexual activity.
Denominator Exclusions: Females who received a pregnancy test to determine contraindications for medication (isotretinoin) or x-ray.

Measure Type: Process
Data Source: Administrative claims, Electronic Clinical Data, Electronic Clinical Data : Imaging/Diagnostic Study, Electronic Clinical Data: Laboratory, Electronic Clinical Data : Pharmacy
Level of Analysis: Health Plan, Integrated Delivery System

IF Endorsement Maintenance – Original Endorsement Date: Aug 10, 2009 **Most Recent Endorsement Date:** May 02, 2012

Maintenance of Endorsement -- Preliminary Analysis

To maintain NQF endorsement endorsed measures are evaluated periodically to ensure that the measures still meets the NQF endorsement criteria ("maintenance"). The emphasis for maintaining endorsement is focused on how effective the measure is for promoting improvements in quality. Endorsed measures should have some experience from the field to inform the evaluation. The emphasis for maintaining endorsement is noted for each criterion.

Criteria 1: Importance to Measure and Report

1a. Evidence
Maintenance measures – less emphasis on evidence unless there is new information or change in evidence since the prior evaluation.

1a. Evidence. The evidence requirements for a *process or intermediate outcome* measure is that it is based on a systematic review (SR) and grading of the body of empirical evidence where the specific focus of the evidence matches what is being measured.

The developer provides the following evidence for this measure:

- **Systematic Review of the evidence specific to this measure?** ☒ **Yes** ☐ **No**

- **Quality, Quantity and Consistency of evidence provided?** ☒ **Yes** ☐ **No**
- **Evidence graded?** ☒ **Yes** ☐ **No**

Evidence Summary or **Summary of prior review in 2012**

The evidence for this measure was based on a 2012 USPSTF recommendation. Measure aligns with USPSTF guideline based on comprehensive meta-analysis -unanimous pass by the prior Committee.

Changes to evidence from last review

☐ **The developer attests that there have been no changes in the evidence since the measure was last evaluated.**

☒ **The developer provided updated evidence for this measure:**

Updates: The developer provided update: USPSTF (2014) recommends screening for chlamydia in sexually active females aged 24 years or younger and in older women who are at increased risk for infection. Grade: B Recommendation (The USPSTF recommends the service. There is high certainty that the net benefit is moderate or there is moderate certainty that the net benefit is moderate to substantial). "Screening for chlamydia has a moderate net benefit in females aged ≤24 y and older women at increased risk for infection."

Evidence synthesis concludes that "Chlamydia screening in young women may reduce pelvic inflammatory disease." USPSTF notes that "the studies it reviewed on the direct effects of screening for chlamydia, including 1 new good-quality RCT, showed mixed results. This led to the change in grade for screening for chlamydia, which is now based on "moderate" certainty of a moderate net benefit rather than "high certainty" of a substantial net benefit."

Exception to evidence- NA

Guidance from the Evidence Algorithm

Process measure (Box 3) → Systematic review with QQC (Box 4) →Moderate certainty that the net benefit is substantial (Box 5b) → moderate

Questions for the Committee:

- *Although the guidelines have been updated, the underlying evidence presented appears to be the directionally the same since the last NQF endorsement review. Does the Committee agree and so there is no need for repeat discussion and vote on Evidence?*

Preliminary rating for evidence: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

1b. Gap in Care/Opportunity for Improvement and 1b. Disparities
Maintenance measures – increased emphasis on gap and variation

1b. Performance Gap. The performance gap requirements include demonstrating quality problems and opportunity for improvement.

The developer presents the following data for 2012 – 2014. Data for 2007 – 2009 was submitted during the prior evaluation. The data recent is further sub-divided into two age groups: 16 -20 years and 21-24 years.

Commercial

Year	2014	2013	2012		2009	2008	2007
N plans	405	410	411		239	251	255
Mean (SD)	45% (10)	44%(10)	44% (10)		43.1 (9.94)	41.7 (10.1)	38.1 (10.7)
MIN-MAX					20.5 – 77.4	16.1 -81.8	15.9 – 92.1
25th and 75th %tile	38 51	37 51	37 50		36.3 49.3	34.8 53.9	31.1 44.3
IQR	13	13	13				

Medicaid

Year	2014	2013	2012		2009	2008	2007
N plans	198	198	176		139	120	130
Mean (SD)	55% (11)	55% (10)	57% (10)		56.7 (10.2)	54.9 (10.3)	50.7 (13.0)
25th and 75th percentiles	40 62	49 63	46 64		50.6 63.7	48.7 61.6	43.3 59.7
IQR	13	14	13				

Disparities
Even though the developer presents evidence from the literature that describes racial/ethnic differences screening rates (higher in African-Americans and Hispanics) and prevalence of the disease (higher in African-Americans and Mexican-Americans) the developer does not collect performance data stratified by race, ethnicity, or language.

Questions for the Committee:

- *This measure has been endorsed by NQF for7 years and used by NCQA even longer. How has performance changed over time?*
- *Is there a gap in care that warrants a national performance measure?*
- *Can this measure be used to address disparities?*

Preliminary rating for opportunity for improvement: ☒ **High** ☐ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 1: Importance to Measure and Report (including 1a, 1b, 1c)

1a. Evidence to Support Measure Focus

Comments:

**Per USPSTF, screening increases rate of chlamydia treatment, which decreases PID and perinatal complications.
I am unclear why the exclusion criteria remove pregnant women. Why does the denominator include sexual activity but the numerator doesn't specify sexual activity. If pap guidelines don't rely on sexual history, why can't we do the same for chlamydia screening?**

1b. Performance Gap

Comments:

Data on disparities not yet available, but is important and it should be.

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability

2a1. Reliability Specifications

Maintenance measures – no change in emphasis – specifications should be evaluated the same as with new measures

2a1. Specifications requires the measure, as specified, to produce consistent (reliable) and credible (valid) results about the quality of care when implemented.

Data source(s): Administrative claims including encounter, lab, pharmacy and radiology claims.

Specifications:

- The level of analysis is health plan or integrated delivery system.
- The denominator includes a claim or encounter indicating sexual activity. Two methods are used to identify sexually active women: pharmacy data for contraceptives and claim/encounter data (see attached: Pregnancy Value Set, Sexual Activity Value Set, and Pregnancy Tests Value Set). A patient only needs to be identified in one method to be eligible for the measure.

- An attached spreadsheet contains numerous ICD-9 and ICD-10 codes for pregnancy, CPT and LOINC codes for chlamydia test, CPT codes for diagnostic radiology, CPT, HCPCS, ICD-9 and 10CM codes for sexual activity, CPT and LOINC codes for pregnancy test
- The developer notes several updates to the codes (HCPCS, LOIN, ICD-9 diagnosis codes) since the prior evaluation.
- A calculation algorithm describes the process of calculating the measure.

Questions for the Committee:

- *Are all the data elements clearly defined? Are all appropriate codes included?*
- *Is the logic or calculation algorithm clear?*
- *Is it likely this measure can be consistently implemented?*

2a2. Reliability Testing Testing attachment
Maintenance measures – less emphasis if no new testing data provided

2a2. Reliability testing demonstrates if the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise enough to distinguish differences in performance across providers.

For maintenance measures, summarize the reliability testing from the prior review:
Signal-to-noise evaluation of the measure score for plans reporting HEDIS data in 2010 --thought to indicate high reliability.

Describe any updates to testing - none

SUMMARY OF TESTING
Reliability testing level ☒ **Measure score** ☐ **Data element** ☐ **Both**
Reliability testing performed with the data source and level of analysis indicated for this measure ☒ **Yes** ☐ **No**

Method(s) of reliability testing
Beta binomial method for assessing signal-to-noise ratio: "A reliability of zero implies that all the variability in a measure is attributable to measurement error. A reliability of one implies that all the variability is attributable to real differences in performance. The higher the reliability score, the greater is the confidence with which one can distinguish the performance of one plan from another." This is an appropriate test for measure score reliability.

Results of reliability testing
Aggregate results of reliability testing for all plans is 0.99 indicating high signal to noise ratio. (We assume this aggregate was based on all the health plans reporting HEDIS data in 2010.)

Guidance from the Reliability Algorithm
Precise specifications (Box 1) → empirical reliability testing (box 2) →used computed performance scores for measure entities (Box 4) → Appropriate method used (Box 5) →High reliability statistic and scope (Box 6a) → High

Questions for the Committee:

- *No updated testing information is presented. The prior testing demonstrated high reliability. Does the Committee think there is a need to re-discuss and re-vote on reliability?*

Preliminary rating for reliability: ☒ **High** ☐ **Moderate** ☐ **Low** ☐ **Insufficient**

2b. Validity
Maintenance measures – less emphasis if no new testing data provided

2b1. Validity: Specifications

2b1. Validity Specifications. This section should determine if the measure specifications are consistent with the evidence.

Specifications consistent with evidence in 1a. ☐ **Yes** ☒ **Somewhat** ☐ **No**

Specification not completely consistent with evidence:

The USPSTF did not recommend a time-frame for screening. This measure expects annual screening.

Question for the Committee:

- *Are the specifications consistent with the evidence?*
- *Is the annual time-frame of the measure based on the evidence?*

2b2. Validity testing

2b2. Validity Testing should demonstrate the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality.

For maintenance measures, summarize the validity testing from the prior review:

Assessment of adequate face validity. The prior Committee did not voice any concerns.

Describe any updates to validity testing - none

SUMMARY OF TESTING

Validity testing level ☒ **Measure score** ☐ **Data element testing against a gold standard** ☐ **Both**

Method of validity testing of the measure score:

- ☒ **Face validity only** (The highest rating possible is moderate)
- ☐ **Empirical validity testing of the measure score**

Validity testing method:

Face validity was assessed using a panel of stakeholders with specific expertise in measurement of women and child health care. This panel included representatives from key stakeholder groups, including experts on women's health, family physicians, health plans, AHRQ and other researchers in the field. The panel reviewed the results of the field test and assessed whether the results were consistent with expectations, whether the measure represented quality care, and whether we were measuring the most important aspect of care in this area.

Validity testing results:

This measure was deemed valid by the developer's expert panel.

Questions for the Committee:

- *No updated testing information is presented. The prior testing demonstrated adequate validity. Does the Committee think there is a need to re-discuss and re-vote on validity?*

2b3-2b7. Threats to Validity

2b3. Exclusions:

- The only exclusion is for females who received a pregnancy test to determine contraindications for medication (isotretinoin) or x-ray.
- The developer does not provide any analysis on the frequency of exclusions.

Questions for the Committee:

- *Are the exclusions consistent with the evidence?*
- *Are any patients or patient groups inappropriately excluded from the measure?*
- *Are the exclusions/exceptions of sufficient frequency and variation across providers to be needed (and outweigh the*

data collection burden)?

2b4. Risk adjustment: **Risk-adjustment method** ☒ **None** ☐ **Statistical model** ☐ **Stratification**

2b5. Meaningful difference *(can statistically significant and clinically/practically meaningful differences in performance measure scores can be identified)*:

- See table under gap. The mean performance result, range, and percentiles describe the distribution of performance among a large group of health plans.

Question for the Committee:

- *Does this measure identify meaningful differences about quality?*

2b6. Comparability of data sources/methods: NA

2b7. Missing Data

- The developers did not address the handling of missing data.

Preliminary rating for validity: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**
In the absence of empirical validity testing, the highest rating possible is moderate.

Committee pre-evaluation comments

Criteria 2: Scientific Acceptability of Measure Properties (including all 2a, 2b, and 2d)

2a1. & 2b1. Specifications
Comments:
Would be a lot easier if all women 16-24 were screened; pregnant women were not excluded.
**Evidence did not recommend a time frame, although this measure expects annual screening. **

2a2. Reliability Testing
Comments:

2b2. Validity Testing
Comments:

2b3. Exclusions Analysis
2b4. Risk Adjustment/Stratification for Outcome or Resource Use Measures
2b5. Identification of Statistically Significant & Meaningful Differences In Performance
2b6. Comparability of Performance Scores When More Than One Set of Specifications
2b7. Missing Data Analysis and Minimizing Bias
Comments:

Criterion 3. Feasibility
Maintenance measures – no change in emphasis – implementation issues may be more prominent

3. Feasibility is the extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

- This measure is based on administrative claims data that is generally considered to be very feasible and low burden.

Questions for the Committee:

- *Does the Committee have any concerns about the feasibility of this measure?*

Preliminary rating for feasibility: ☒ **High** ☐ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments
Criteria 3: Feasibility

3a. Byproduct of Care Processes
3b. Electronic Sources
3c. Data Collection Strategy
Comments:
I think it would be a lot easier to do for all women 16-24.

Criterion 4: Usability and Use
Maintenance measures – increased emphasis – much greater focus on measure use and usefulness, including both impact /improvement and unintended consequences

4. Usability and Use evaluate the extent to which audiences (e.g., consumers, purchasers, providers, policymakers) use or could use performance results for both accountability and performance improvement activities.

Current uses of the measure
Publicly reported? ☒ **Yes** ☐ **No**

Current use in an accountability program? ☒ **Yes** ☐ **No**
OR
Planned use in an accountability program? ☐ **Yes** ☐ **No**

Accountability program details
The developer reports on multiple uses of the measures including an eMeasure version and clinician level measure that are not included NQF's endorsement.

- Performance results for the measure are available for a fee to licensees of NCQA's Quality Compass.
- This measure is publicly reported by medical groups in California by the state Office of the Patient Advocate. See results for groups in Los Angeles (You can select other counties in California to view their results.)
- Many individual organizations publish their results of this measure such as Premera, ConnectiCare, Aetna Health, Inc. (Georgia).
- CALIFORNIA'S VALUE BASED PAY FOR PERFORMANCE PROGRAM: This measure is used in the California P4P program, which is the largest non-governmental physician incentive program in the United States.

Improvement results The data over the years 2007-2014 show slight improvement over time.

Unexpected findings (positive or negative) during implementation
The developer reports that "there were no unexpected findings or unexpected benefits."

Potential harms None have been identified.

Feedback:

- No feedback submitted via QPS

***Questions for the Committee**:*

- *This measure has been in use for many years – what has been the impact?*
- *How do you interpret performance results in which overall performance seems low, yet performance rates in the most recent 3 years are unchanged?*
- *Is the Committee aware of any unintended consequences or potential harms from the measure? Any unexpected benefits from use of the measure?*

○ *Do the benefits of the measure outweigh any potential unintended consequences?*

Preliminary rating for usability and use: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 4: Usability and Use

4a. Accountability and Transparency
4b. Improvement
4c. Unintended Consequences
Comments:
It's concerning that the performance rates have not changed in three years.

Criterion 5: Related and Competing Measures

Related or competing measures
None

Harmonization -NA

Pre-meeting public and member comments

- This measure permits an exclusion if an x-ray procedure is performed within 7 days of a pregnancy test. The allowed x-ray procedures are defined in CMS eCQM 153 as a test contained in the grouper: Diagnostic Study, Order: X-Ray Study (all inclusive)" using "X-Ray Study (all inclusive) Grouping Value Set (2.16.840.1.113883.3.464.1003.198.12.1034). This is a LOINC code grouper. Our EHR (Epic) and the procedure database (AMA) only link CPT codes to our x-ray procedures, not LOINC. Is there a procedure grouper based on CPT codes instead of LOINC that can be used for this measure? If not, what are recommended next steps for setting up this measure for MU reporting?

- The measure seems very reasonable. The denominator exclusions could be defined to account for the patients who may be taking birth control pills for health-related reasons but who are not sexually active. Please at least remove all patients with pregnancy tests due to surgery within a week as well as those receiving X-rays and Accutane. Consider removing some menstrual diagnoses such as PCOS, ovarian failure, amenorrhea, etc. from the denominator. A number of these need hormonal treatment and are not suggestive of sexual activity. We also recommend adding screenings at behavioral health clinics in order to capture populations at the highest risk for chlamydia.

1. IMPACT, OPPORTUITY, EVIDENCE - IMPORTANCE TO MEASURE AND REPORT

Importance to Measure and Report is a threshold criterion that must be met in order to recommend a measure for endorsement. All three subcriteria must be met to pass this criterion. See guidance on evidence.

Measures must be judged to be important to measure and report in order to be evaluated against the remaining criteria. **(evaluation criteria)**

1c.1 **Structure-Process-Outcome Relationship** *(Briefly state the measure focus, e.g., health outcome, intermediate clinical outcome, process, structure; then identify the appropriate links, e.g., structure-process-health outcome; process-health outcome; intermediate clinical outcome-health outcome)***:**

There is good evidence that screening for Chlamydial infection in women who are at increased risk can reduce the incidence of pelvic inflammatory disease (PID), infertility and perinatal infections. The US Preventive Services Task Force (USPSTF) concluded that the benefits of screening women at increased risk are substantial.

1c.2-3 **Type of Evidence** *(Check all that apply)***:**

Clinical Practice Guideline

1c.4 **Directness of Evidence to the Specified Measure** *(State the central topic, population, and outcomes addressed in the body of evidence and identify any differences from the measure focus and measure target population)***:**

Nonpregnant women: To date there has been one good randomized controlled trial that indicates screening reduces pelvic inflammatory disease in non-pregnant women.

Pregnant women: The previous Task Force recommendations for screening pregnant women were based on two studies that demonstrated improved pregnancy outcomes following treatment of chlamydial infection. In a time-series design study, untreated patients had a significantly higher incidence of premature rupture of membranes and low birth weight as well as a lower infant survival rate compared to treated patients and patients with negative cultures. In a case-control study, the frequencies of premature rupture of membranes, premature contractions, and small-for-gestational-age infants were significantly lower among successfully treated patients compared to chlamydia-positive patients who were unresponsive to treatment, but they were not significantly different when compared to chlamydia-negative control patients. No other studies were identified

Men: No studies were found that described the effectiveness of screening or early treatment for men in reducing transmission to women or of preventing acute infections or complications in men. Many investigators advocate screening men as the next essential step to reduce infections, complications, and recurrences in women, as well as to improve the health of men themselves. However, these health outcomes have not yet been studied.

No adequately controlled study has prospectively addressed whether screening reduces the prevalence of infection, although several studies have been published that report declining prevalence rates in women after instituting chlamydia testing and treatment programs. Changes in population prevalence rates have not been well documented because few studies have employed a representative population sample. Other unmeasured factors, such as condom use, changes in sexual behavior, and changes in testing methods, could also be responsible for changes in prevalence rates.

AHRQ. Screening for Chlamydial Infection: Summary of the Evidence. http://archive.ahrq.gov/clinic/ajpmsuppl/nelson1.htm#section6 (July 10, 2011).

1c.5 **Quantity of Studies in the Body of Evidence** *(Total number of studies, not articles)***:** The measure is based on a USPSTF guideline that is based on a comprehensive meta-analysis (see USPSTF report for full number of studies)

1c.6 **Quality of Body of Evidence** (*Summarize the certainty or confidence in the estimates of benefits and harms to patients across studies in the body of evidence resulting from study factors. Please address: a) study design/flaws; b) directness/indirectness of the evidence to this measure (e.g., interventions, comparisons, outcomes assessed, population included in the evidence); and c) imprecision/wide confidence intervals due to few patients or events):* High.

Non-pregnant women at increased risk. There is good evidence that screening for Chlamydial infection in women who are at increased risk can reduce the incidence of pelvic inflammatory disease (PID). The US Preventive Services Task Force (USPSTF) concluded that the benefits of screening women at increased risk are substantial.

Pregnant women at increased risk. There are no studies evaluating the effectiveness of screening for chlamydial infection in pregnant women who are at increased risk. The USPSTF, however, found the following: 1) screening identifies infection in asymptomatic pregnant women; 2) there is a relatively high prevalence of infection among pregnant women who are at increased risk; and 3) there is fair evidence of improved pregnancy and birth outcomes for women who are treated for chlamydial infection. The USPSTF concluded that the benefits of screening pregnant women who are at increased risk are substantial.

Women not at increased risk. The USPSTF identified no studies documenting the benefits of screening women, including pregnant women, who are not at increased risk for chlamydial infection. While recognizing the potential benefit to women identified through screening, the USPSTF concluded the overall benefit of screening would be small, given the low prevalence of infection among women not at increased risk.

Men. While concluding that the direct benefit to men of screening was likely to be small, the USPSTF noted that screening for chlamydial infection in men may be beneficial if it were to lead to a decreased incidence of chlamydial infection in women. The USPSTF did not, however, find evidence to support this outcome, and therefore concluded that the benefits of screening men are unknown. The USPSTF identified this as a critical gap in the evidence.

1c.7 **Consistency of Results across Studies** *(Summarize the consistency of the magnitude and direction of the effect):* Consistent

1c.8 **Net Benefit** *(Provide estimates of effect for benefit/outcome; identify harms addressed and estimates of effect; and net benefit - benefit over harms)***:**

The USPSTF determined there was a positive net benefit

1c.9 **Grading of Strength/Quality of the Body of Evidence**. Has the body of evidence been graded? Yes

1c.10 **If body of evidence graded, identify the entity that graded the evidence including balance of representation and any disclosures regarding bias:** USPSTF

1c.11 **System Used for Grading the Body of Evidence:** USPSTF

1c.12 **If other, identify and describe the grading scale with definitions:**

1c.13 **Grade Assigned to the Body of Evidence:**

1c.14 **Summary of Controversy/Contradictory Evidence:** While some research suggests it would be prudent to screen males for chlamydia, at present, the U.S. Preventive Services Task Force does not recommend screening males.

1c.15 **Citations for Evidence other than Guidelines*(Guidelines addressed below)*:**

Voelker, R. (2010). Experts Reconsider Wisdom of Limiting Chlamydia Screening Only to Women. JAMA 303(9):823-824.

1c.16 Quote verbatim, the specific guideline recommendation *(Including guideline # and/or page #)*:

The USPSTF (2014) recommends screening for chlamydia in sexually active females aged 24 years or younger and in older women who are at increased risk for infection. Grade: B Recommendation.

The US Preventive Services Task Force (USPSTF) recommends screening for chlamydial infection for all sexually active non-pregnant young women aged 24 and younger, and for older non-pregnant women who are at increased risk. A recommendation

The USPSTF recommends screening for chlamydial infection for all pregnant women aged 24 and younger, and for older pregnant women who are at increased risk. B recommendation

The USPSTF recommends against routinely providing screening for chlamydial infection for women aged 25 and older, whether or not they are pregnant, if they are not at increased risk. C recommendation

The USPSTF concludes that the current evidence is insufficient to assess the balance of benefits and harms of screening for chlamydial infection for men. I statement.

CDC recommends screening all sexually active women aged 25 years and younger and older women with risk factors (e.g., those who have a new sex partner or multiple sex partners).

All pregnant women should be routinely tested at the first prenatal visit. Pregnant women aged 25 years and younger and those at increased risk should be re-tested during the third trimester to prevent maternal postnatal complications and chlamydial infection in the infant.

1c.17 Clinical Practice Guideline Citation:

LeFevre, M.L. Screening for Chlamydia and Gonorrhea: U.S. Preventive Services Task Force Recommendation Statement. Ann Intern Med. 2014;161(12):902-10.

Guideline available from: http://www.uspreventiveservicestaskforce.org/Page/Document/RecommendationStatementFinal/chlamydia-and-gonorrhea-screening, accessed February 16, 2016.

1. U.S. Preventive Services Task Force. Screening for chlamydial infection: U.S. Preventive Services Task Force recommendation statement. Ann Intern Med 2007 Jul 17;147(2):128-34.

2. Centers for Disease Control and Prevention (CDC). http://www.guideline.gov/content.aspx?id=9672.

1c.18 National Guideline Clearinghouse or other URL: http://www.guideline.gov/content.aspx?id=10408#Section424

1c.19 Grading of Strength of Guideline Recommendation. Has the recommendation been graded? No

1c.20 If guideline recommendation graded, identify the entity that graded the evidence including balance of representation and any disclosures regarding bias:

1c.21 System Used for Grading the Strength of Guideline Recommendation: USPSTF

1c.22 If other, identify and describe the grading scale with definitions:

1c.23 Grade Assigned to the Recommendation:

1c.24 Rationale for Using this Guideline Over Others: This measure aligns with the USPSTF guidelines, which is the gold standard for evidence reviews on preventive services. In addition, NCQA convened an expert panel of diverse stakeholders to review the guidelines and evidence for this measure. The panel determined the measure was scientifically sound using the full body of evidence and guidelines for this measure concept.

Based on the NQF descriptions for rating the evidence, what was the <u>developer's assessment</u> of the quantity, quality, and consistency of the body of evidence?

1c.25 Quantity: High **1c.26** Quality: High**1c.27** Consistency**:** High

1. Evidence, Performance Gap, Priority – Importance to Measure and Report

Extent to which the specific measure focus is evidence-based, important to making significant gains in healthcare quality, and improving health outcomes for a specific high-priority (high-impact) aspect of healthcare where there is variation in or overall less-than-optimal performance. ***Measures must be judged to meet all subcriteria to pass this criterion and be evaluated against the remaining criteria.***

1a. Evidence to Support the Measure Focus – See attached Evidence Submission Form
0033_CHL_Evidence.doc

1b. Performance Gap

Demonstration of quality problems and opportunity for improvement, i.e., data demonstrating:

- considerable variation, or overall less-than-optimal performance, in the quality of care across providers; and/or
- disparities in care across population groups.

1b.1. **Briefly explain the rationale for this measure** *(e.g., the benefits or improvements in quality envisioned by use of this measure)*
This measure assesses the percentage women 16-24 years of age who were identified as sexually active and who received a test for chlamydia. The improvement in quality envisioned by the use of this measure is increased identification of untreated chlamydia infections in women that can lead to serious and irreversible complications, and can be unknowingly transmitted to sexual partners. Despite the availability of effective treatments, a large proportion of sexually active individuals continue to go undiagnosed due to the disease's asymptomatic nature. Early detection, screening, and treatment have proven to be effective in managing and preventing chlamydia.

1b.2. **Provide performance scores on the measure as specified (current and over time) at the specified level of analysis.** *(This is required for endorsement maintenance. Include mean, std dev, min, max, interquartile range, scores by decile. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included). This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*
The following data are extracted from HEDIS data collection reflecting the most recent years of measurement for this measure. Performance data are summarized at the health plan level by mean, standard deviation, and performance at the 10th, 25th, 50th, 75th and 90th percentile. Data are stratified by year and product line (i.e. commercial and Medicaid).

Chlamydia Screening in Women

Commercial Rate – 16 to 20 years of age

YEAR| MEAN | ST DEV | 10TH | 25TH | 50TH | 75TH | 90TH | Interquartile Range

2014 | 40% | 11% | 29% | 33% | 38% | 45% | 56% | 12

2013 | 40% | 10% | 29% | 33% | 38% | 45% | 54% | 12

2012 | 40% | 10% | 30% | 34% | 38% | 45% | 53% | 12

Commercial Rate – 21 to 24 years of age

YEAR| MEAN | ST DEV | 10TH | 25TH | 50TH | 75TH | 90TH | Interquartile Range

2014 | 49% | 10% | 38% | 43% | 47% | 56% | 63% | 13

2013 | 48% | 10% | 36% | 41% | 46% | 55% | 63% | 14

2012 | 47% | 10% | 34% | 40% | 46% | 54% | 62% | 14

Commercial Rate – Total

YEAR| MEAN | ST DEV | 10TH | 25TH | 50TH | 75TH | 90TH | Interquartile Range

2014 | 45% | 10% | 34% | 38% | 43% | 51% | 60% | 13

2013 | 44% | 10% | 33% | 37% | 43% | 51% | 58% | 13

2012 | 44% | 10% | 32% | 37% | 42% | 50% | 57% | 13

Medicaid Rate – 16 to 20 years of age

YEAR| MEAN | ST DEV | 10TH | 25TH | 50TH | 75TH | 90TH | Interquartile Range

2014 | 51% | 11% | 37% | 44% | 50% | 58% | 67% | 13

2013 | 51% | 11% | 37% | 44% | 52% | 59% | 64% | 15

2012 | 53% | 10% | 41% | 47% | 54% | 59% | 66% | 13

Medicaid Rate – 21 to 24 years of age

YEAR| MEAN | ST DEV | 10TH | 25TH | 50TH | 75TH | 90TH | Interquartile Range

2014 | 60% | 10% | 47% | 54% | 61% | 67% | 72% | 13

2013 | 62% | 10% | 49% | 57% | 63% | 69% | 72% | 13

2012 | 63% | 9% | 52% | 59% | 65% | 71% | 73% | 12

Medicaid Rate – Total

YEAR| MEAN | ST DEV | 10TH | 25TH | 50TH | 75TH | 90TH | Interquartile Range

2014 | 55% | 11% | 40% | 49% | 54% | 62% | 69% | 13

2013 | 55% | 10% | 41% | 49% | 55% | 63% | 67% | 14

2012 | 57% | 10% | 46% | 51% | 57% | 64% | 69% | 13

The data references are extracted from HEDIS data collection reflecting the most recent years of measurement for this measure. In 2013, HEDIS measures covered more than 171 million people from 814 HMOs and 353 PPOs. Below is a description of the denominator for this measure. It includes the number of health plans included in HEDIS data collection and the median eligible population for the measure across health plans.

Commercial – 16 to 20 years of age

YEAR | N Plans | Median Denominator Size per plan

2014 | 401 | 1323

2013 | 409 | 1390

2012 | 411 | 1376

Commercial – 21 to 24 years of age

YEAR | N Plans | Median Denominator Size per plan

2014 | 403 | 1561

2013 | 410 | 1590

2012 | 411 | 1549

Commercial – Total

YEAR | N Plans | Median Denominator Size per plan

2014 | 405 | 2922

2013 | 412 | 2984

2012 | 415 | 2956

Medicaid – 16 to 20 years of age

YEAR | N Plans | Median Denominator Size per plan

2014 | 189 | 1979

2013 | 191 | 1556

2012 | 171 | 1655

Medicaid – 21 to 24 years of age

YEAR | N Plans | Median Denominator Size per plan

2014 | 195 | 1237

2013 | 190 | 957

2012 | 172 | 1034

Medicaid – Total

YEAR | N Plans | Median Denominator Size per plan

2014 | 198 | 3082

2013 | 198 | 2362

2012 | 176 | 2508

1b.3. If no or limited performance data on the measure as specified is reported in 1b2, then provide a summary of data from the literature that indicates opportunity for improvement or overall less than optimal performance on the specific focus of measurement.
NA

1b.4. (11)Provide disparities data from the measure as specified (current and over time) by population group, e.g., by race/ethnicity, gender, age, insurance status, socioeconomic status, and/or disability. *(This is required for endorsement*

maintenance. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities include.) This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.

HEDIS data, including data for this measure, are stratified by type of insurance (i.e., Commercial, Medicaid, Medicare). NCQA does not currently collect performance data stratified by race, ethnicity, or language. Escarce et al. have described in detail the difficulty of collecting valid data on race, ethnicity and language at the health plan level (Escarce 2011). While not specified in the measure, this measure can also be stratified by demographic variables, such as race/ethnicity or socioeconomic status, in order to assess the presence of health care disparities. The HEDIS Health Plan Measure Set contains two measures that can assist with stratification to assess health care disparities. The Race/Ethnicity Diversity of Membership and the Language Diversity of Membership measures were designed to promote standardized methods for collecting these data and follow Office of Management and Budget and Institute of Medicine guidelines for collecting and categorizing race/ethnicity and language data. In addition, NCQA's Multicultural Health Care Distinction Program outlines standards for collecting, storing and using race/ethnicity and language data to assess health care disparities. Based on extensive work by NCQA to understand how to promote culturally and linguistically appropriate services among plans and providers, we have many examples of how health plans have used HEDIS measures to design quality improvement programs to decrease disparities in care.

Escarce, J.J., R. Carreon, G. Vesolovskiy, E.H. Lawson. 2011. "Collection Of Race And Ethnicity Data By Health Plans Has Grown Substantially, But Opportunities Remain To Expand Efforts." Health Affairs 30(10): 1984-1991. doi: 10.1377/hlthaff.2010.1117.

1b.5. **If no or limited data on disparities from the measure as specified is reported in 1b4, then provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement. Include citations.**

Studies suggest racial/ethnic disparities in chlamydia infection exist. A recent study found that African-American and Hispanic women are screened at higher rates than their Caucasian counterparts. African-American women were three times more likely to be tested, and Hispanic women nearly 13 times more likely compared to Caucasian women (Wiehe 2011). Given differential rates in screening racial/ethnic groups, varying prevalence according to race/ethnicity exist. According to recent data, prevalence among non-Hispanic African-Americans is approximately seven times higher than the prevalence among non-Hispanic whites, and prevalence among Mexican-Americans is approximately three times the prevalence among non-Hispanic whites. Among sexually active females aged 14-24 years, approximately one in seven non-Hispanic black females are infected with chlamydia; one in 22 Mexican-American females, and one in 55 non-Hispanic white females (Torrone 2014).

Torrone, E., J. Papp, H. Weinstock. 2014. "Prevalence of Chlamydia trachomatis Genital Infection Among Persons Aged 14-39 Years – United States, 2007-2012." MMWR Morb Mortal Wkly Rep. 63(38):834-8.

Wiehe, S.E., M.B. Rosenman, J. Wang, B.P. Katz, J.D. Fortenberry. 2011. "Chlamydia screening among young women: individual- and provider-level differences in testing." Pediatrics. 127(2):e336-44. doi: 10.1542/peds.2010-0967.

1c. High Priority (previously referred to as High Impact)

The measure addresses:

- a specific national health goal/priority identified by DHHS or the National Priorities Partnership convened by NQF; OR
- a demonstrated high-priority (high-impact) aspect of healthcare (e.g., affects large numbers of patients and/or has a substantial impact for a smaller population; leading cause of morbidity/mortality; high resource use (current and/or future); severity of illness; and severity of patient/societal consequences of poor quality).

1c.1. **Demonstrated high priority aspect of healthcare**

Affects large numbers, A leading cause of morbidity/mortality, High resource use, Patient/societal consequences of poor quality, Severity of illness

1c.2. **If Other:**

1c.3. **Provide epidemiologic or resource use data that demonstrates the measure addresses a high priority aspect of healthcare. List citations in 1c.4.**

Chlamydia trachomatis is the most common sexually transmitted bacterial infection in the US. It is estimated that 1 in 20 sexually active young women aged 14-24 years has chlamydia (Torrone 2014). It is often known as a "silent" disease because most infected people have no symptoms and therefore are unaware they have an infection. Although Chlamydia symptoms are usually mild or nonexistent, untreated infections can lead to serious and irreversible complications. Among women with chlamydial infection,

symptomatic pelvic inflammatory disorder occurs in about 10 to 15 percent of women with untreated chlamydia, which can cause permanent damage to the fallopian tubes, uterus, and surrounding tissues and lead to chronic pelvic pain, tubal factor infertility, and potentially fatal ectopic pregnancy (CDC 2014). Pregnant women infected with untreated chlamydia are susceptible to pre-term delivery, as well as ophthalmia neonatorum (conjunctivitis) and pneumonia in newborns (Rours 2011). Over 1.4 million chlamydial infections were reported to the Centers for Disease Control and Prevention from 50 states and the District of Columbia in 2012. Since many cases are not reported or even diagnosed, it is estimated that there are actually 2.86 million new cases of chlamydia each year (CDC 2014).

1c.4. **Citations for data demonstrating high priority provided in 1a.3**

Centers for Disease and Prevention. 2014. "Chlamydia – CDC Fact Sheet." http://www.cdc.gov/std/chlamydia/stdfact-chlamydia-detailed.htm_ (Feb. 12, 2016).

Centers for Disease Control and Prevention. 2012. "Sexually Transmitted Diseases Surveillance." Atlanta, GA: Centers for Disease Control and Prevention; 2014. www.cdc.gov/std/stats12/default.htm (Feb. 12, 2016).

Rours, G.I., L. Duijts, H.A. Moll, et al. 2011. "Chlamydia trachomatis infection during pregnancy associated with preterm delivery: a population-based prospective cohort study." Eur J Epidemiol. 26(6):493-502. doi: 10.1007/s10654-011-9586-1.

Torrone, E., J. Papp, H. Weinstock. 2014. "Prevalence of Chlamydia trachomatis Genital Infection Among Persons Aged 14-39 Years – United States, 2007-2012." MMWR Morb Mortal Wkly Rep. 63(38):834-8.

1c.5. **If a PRO-PM (e.g. HRQoL/functional status, symptom/burden, experience with care, health-related behaviors), provide evidence that the target population values the measured PRO and finds it meaningful.** *(Describe how and from whom their input was obtained.)*

NA

2. Reliability and Validity—Scientific Acceptability of Measure Properties

Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. ***Measures must be judged to meet the subcriteria for both reliability and validity to pass this criterion and be evaluated against the remaining criteria.***

2a.1. Specifications The measure is well defined and precisely specified so it can be implemented consistently within and across organizations and allows for comparability. eMeasures should be specified in the Health Quality Measures Format (HQMF) and the Quality Data Model (QDM).

De.5. **Subject/Topic Area** *(check all the areas that apply):*
Infectious Diseases : Sexually Transmitted, Perinatal and Reproductive Health : Screening, Prevention : Screening

De.6. **Cross Cutting Areas** *(check all the areas that apply):*
Prevention : Screening

S.1. **Measure-specific Web Page** *(Provide a URL link to a web page specific for this measure that contains current detailed specifications including code lists, risk model details, and supplemental materials. Do not enter a URL linking to a home page or to general information.)*
NA

S.2a. **If this is an eMeasure**, HQMF specifications must be attached. Attach the zipped output from the eMeasure authoring tool (MAT) - if the MAT was not used, contact staff. (Use the specification fields in this online form for the plain-language description of the specifications)
This is not an eMeasure **Attachment:**

S.2b. **Data Dictionary, Code Table, or Value Sets** *(and risk model codes and coefficients when applicable) must be attached. (Excel or csv file in the suggested format preferred - if not, contact staff)*
Attachment **Attachment:** 0033_CHL_Value_Sets.xlsx

S.3. **For endorsement maintenance**, please briefly describe any changes to the measure specifications since last endorsement date and explain the reasons.
As part of NCQA's annual measure maintenance, we routinely make coding and other specification tweaks to ensure the measure remains up-to-date with current practice and based on feedback received from measure useres. Below are specific relevant changes we have made to the measure since its last endorsement in 2012.

* Added HCPCS code G0450 to Table CHL-B.

* Added ICD-9-CM Diagnosis codes 302.76, 625.0 to Table CHL-B.

* Added LOINC code 69002-4 to Table CHL-B.

S.4. **Numerator Statement** *(Brief, narrative description of the measure focus or what is being measured about the target population, i.e., cases from the target population with the target process, condition, event, or outcome)*
IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
Females who were tested for chlamydia during the measurement year.

S.5. **Time Period for Data** *(What is the time period in which data will be aggregated for the measure, e.g., 12 mo, 3 years, look back to August for flu vaccination? Note if there are different time periods for the numerator and denominator.)*
Numerator: 12 months

Denominator: 12 months

Exclusions: 12 months plus 6 days

S.6. **Numerator Details** *(All information required to identify and calculate the cases from the target population with the target process, condition, event, or outcome such as definitions, specific data collection items/responses, code/value sets – Note: lists of*

individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)
IF an OUTCOME MEASURE, describe how the observed outcome is identified/counted. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
Females who had at least one test for chlamydia (see attached: Chlamydia Tests Value Set) during the measurement year.

S.7. Denominator Statement *(Brief, narrative description of the target population being measured)*
Females 16-24 years who had a claim or encounter indicating sexual activity.

S.8. Target Population Category *(Check all the populations for which the measure is specified and tested if any):*
Children's Health, Maternal Health

S.9. Denominator Details *(All information required to identify and calculate the target population/denominator such as definitions, specific data collection items/responses , code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
All female patients 16-24 years as of December 31 of the measurement year and who were identified as sexually active during the measurement year.

Sexually active: Two methods are used to identify sexually active women: pharmacy data (see CHL-A: Prescriptions to Identify Contraceptives) and claim/encounter data (see attached: Pregnancy Value Set, Sexual Activity Value Set, and Pregnancy Tests Value Set). Both methods are used to identify the eligible population; however, a patient only needs to be identified in one method to be eligible for the measure.

Table CHL-A: Prescriptions to Identify Contraceptives

--Contraceptives: Desogestrel-ethinyl estradiol; Dienogest-estradiol multiphasic; Drospirenone-ethinyl estradiol; Drospirenone-ethinyl estradiol-levomefolate biphasic; Ethinyl estradiol-ethynodiol; Ethinyl estradiol-etonogestrel; Ethinyl estradiol-levonorgestrel; Ethinyl estradiol-norelgestromin; Ethinyl estradiol-norethindrone; Ethinyl estradiol-norgestimate; Ethinyl estradiol-norgestrel; Etonogestrel; Levonorgestrel; Medroxyprogesterone; Mestranol-norethindrone; Norethindrone
--Diaphragm
--Spermicide: Nonxynol 9

S.10. Denominator Exclusions *(Brief narrative description of exclusions from the target population)*
Females who received a pregnancy test to determine contraindications for medication (isotretinoin) or x-ray.

S.11. Denominator Exclusion Details *(All information required to identify and calculate exclusions from the denominator such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
Exclude members from the denominator who were identified as sexually active based on a pregnancy test alone (see attached: Pregnancy Tests Value Set) AND who meet either of the following:

1) A pregnancy test (see attached: Pregnancy Test Exclusion Value Set) during the measurement year AND a prescription for isotretinoin (see Table CHL-E: Medications to Identify Exclusions) on the date of the pregnancy test or the 6 days after the pregnancy test.

2) A pregnancy test (see attached: Pregnancy Test Exclusion Value Set) during the measurement year AND a x-ray (see attached: Diagnostic Radiology Value Set) on the date of the pregnancy test or the 6 days after the pregnancy test.

Table CHL-E: Medications to Identify Exclusions

Retinoid: Isotretinoin

S.12. Stratification Details/Variables *(All information required to stratify the measure results including the stratification variables, definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format with at S.2b)*
The measure includes two age stratifications and a total rate:

1) 16-20 years.

2) 21-24 years.

3) Total

S.13. Risk Adjustment Type (Select type. Provide specifications for risk stratification in S.12 and for statistical model in S.14-15)
No risk adjustment or risk stratification
If other:

S.14. Identify the statistical risk model method and variables *(Name the statistical method - e.g., logistic regression and list all the risk factor variables. Note - risk model development and testing should be addressed with measure testing under Scientific Acceptability)*
NA

S.15. Detailed risk model specifications *(must be in attached data dictionary/code list Excel or csv file. Also indicate if available at measure-specific URL identified in S.1.)*
Note: Risk model details (including coefficients, equations, codes with descriptors, definitions), should be provided on a separate worksheet in the suggested format in the Excel or csv file with data dictionary/code lists at S.2b.

S.15a. Detailed risk model specifications (*if not provided in excel or csv file at S.2b*)
NA

S.16. Type of score:
Rate/proportion
If other:

S.17. Interpretation of Score *(Classifies interpretation of score according to whether better quality is associated with a higher score, a lower score, a score falling within a defined interval, or a passing score)*
Better quality = Higher score

S.18. Calculation Algorithm/Measure Logic *(Describe the calculation of the measure score as an ordered sequence of steps including identifying the target population; exclusions; cases meeting the target process, condition, event, or outcome; aggregating data; risk adjustment; etc.)*
Refer to items S.9 (Denominator details) and S.2b (Data Dictionary) for tables.

Step 1 Determine the eligible population. To do so, identify all female patients in the specified age range who had a claim/encounter indicating sexual activity (Pregnancy Value Set, Sexual Activity Value Set, Pregnancy Tests Value Set) and/or were dispensed prescription contraceptives (Table CHL-A) during the measurement year.

Step 2 Exclude patients who qualified for the eligible population based on a pregnancy test (Pregnancy Tests Value Set) alone AND who meet either of the following: (1) A pregnancy test (Pregnancy Test Exclusion Value Set) during the measurement year AND a prescription for isotretinoin (Table CHL-E) on the date of the pregnancy test or the 6 days after the pregnancy test, (2) A pregnancy test (Pregnancy Test Exclusion Value Set) during the measurement year AND an x-ray (Diagnostic Radiology Value Set) on the date of the pregnancy test or the 6 days after the pregnancy test.

Step 3 Determine the numerator. Determine the number of patients in the remaining eligible population who had at least one chlamydia test (Chlamydia Tests Value Set) during the measurement year.

Step 4 Report two age stratifications (16-20 years and 21-24 years), and a total rate. The total is the sum of the age stratifications.

S.19. Calculation Algorithm/Measure Logic Diagram URL or Attachment *(You also may provide a diagram of the Calculation Algorithm/Measure Logic described above at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
No diagram provided

S.20. Sampling *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*

IF a PRO-PM, identify whether (and how) proxy responses are allowed.
NA

S.21. **Survey/Patient-reported data** *(If measure is based on a survey, provide instructions for conducting the survey and guidance on minimum response rate.)*
IF a PRO-PM, specify calculation of response rates to be reported with performance measure results.
NA

S.22. **Missing data** (specify how missing data are handled, e.g., imputation, delete case.)
Required for Composites and PRO-PMs.
NA

S.23. **Data Source** *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*
If other, please describe in S.24.
Administrative claims, Electronic Clinical Data, Electronic Clinical Data : Imaging/Diagnostic Study, Electronic Clinical Data : Laboratory, Electronic Clinical Data : Pharmacy

S.24. **Data Source or Collection Instrument** *(Identify the specific data source/data collection instrument e.g. name of database, clinical registry, collection instrument, etc.)*
IF a PRO-PM, identify the specific PROM(s); and standard methods, modes, and languages of administration.
This measure is based on administrative claims collected in the course of providing care to health plan members. NCQA collects the Healthcare Effectiveness Data and Information Set (HEDIS) data for this measure directly from Health Management Organizations and Preferred Provider Organizations via NCQA's online data submission system.

S.25. **Data Source or Collection Instrument** *(available at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
No data collection instrument provided

S.26. **Level of Analysis** *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*
Health Plan, Integrated Delivery System

S.27. **Care Setting** *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*
Ambulatory Care : Clinician Office/Clinic
If other:

S.28. **COMPOSITE Performance Measure** - Additional Specifications *(Use this section as needed for aggregation and weighting rules, or calculation of individual performance measures if not individually endorsed.)*
NA

2a. Reliability – See attached Measure Testing Submission Form
2b. Validity – See attached Measure Testing Submission Form
0033_MeasureTesting_MSF5.0_Data.doc

2. RELIABILITY & VALIDITY - SCIENTIFIC ACCEPTABILITY OF MEASURE PROPERTIES

Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. **(evaluation criteria)**

Measure testing must demonstrate adequate reliability and validity in order to be recommended for endorsement. Testing may be conducted for data elements and/or the computed measure score. Testing information and results should be entered in the appropriate field. Supplemental materials may be referenced or attached in item 2.1. See guidance on measure testing.

2a2. Reliability Testing. (*Reliability testing was conducted with appropriate method, scope, and adequate demonstration of reliability.*)

2a2.1 Data/Sample *(Description of the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)***:**

HEDIS Health Plan performance data from 2010

2a2.2 Analytic Method *(Describe method of reliability testing & rationale)***:**

Reliability was estimated by using the beta-binomial model. Beta-binomial is a better fit when estimating the reliability of simple pass/fail rate measures as is the case with most HEDIS® health plan measures. The beta-binomial model assumes the plan score is a binomial random variable conditional on the plan´s true value that comes from the beta distribution. The beta distribution is usually defined by two parameters, alpha and beta. Alpha and beta can be thought of as intermediate calculations to get to the needed variance estimates. The beta distribution can be symmetric, skewed or even U-shaped.

Reliability used here is the ratio of signal to noise. The signal in this case is the proportion of the variability in measured performance that can be explained by real differences in performance. A reliability of zero implies that all the variability in a measure is attributable to measurement error. A reliability of one implies that all the variability is attributable to real differences in performance. The higher the reliability score, the greater is the confidence with which one can distinguish the performance of one plan from another. A reliability score greater than or equal to 0.7 is considered very good.

2a2.3 Testing Results *(Reliability statistics, assessment of adequacy in the context of norms for the test conducted)***:**

Reliability statistic for Chlamydia screening is 0.99.

2b. VALIDITY. Validity, Testing, including all Threats to Validity: H☐ M☐ L☐ I ☐

2b1.1 Describe how the measure specifications *(measure focus, target population, and exclusions)* **are consistent with the evidence cited in support of the measure focus (***criterion 1c)* **and identify any differences from the evidence:**

The measure is aligned with current guidelines

2b2. Validity Testing. (*Validity testing was conducted with appropriate method, scope, and adequate demonstration of validity.)*

2b2.1 Data/Sample *(Description of the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)***:**

Field test data

2b2.2 Analytic Method *(Describe method of validity testing and rationale; if face validity, describe systematic assessment)***:**

NCQA tested the measure for face validity using a panel of stakeholders with specific expertise in measurement of women and child health care. This panel included representatives from key stakeholder groups, including experts on women's health, family physicians, health plans, AHRQ and other researchers in the field. (See list of members of Women & Child Measurement Advisory Panel (WCMAP)). Experts reviewed the results of the field test and assessed whether the results were consistent with expectations, whether the measure represented quality care, and whether we were measuring the most important aspect of care in this area.

2b2.3 Testing Results *(Statistical results, assessment of adequacy in the context of norms for the test conducted; if face validity, describe results of systematic assessment)***:**

This measure was deemed valid by the expert panel.

POTENTIAL THREATS TO VALIDITY. (*All potential threats to validity were appropriately tested with adequate results.*)

2b3. Measure Exclusions. (*Exclusions were supported by the clinical evidence in 1c or appropriately tested with results demonstrating the need to specify them.*)

2b3.1 Data/Sample for analysis of exclusions *(Description of the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)***:**

2b3.2 Analytic Method (*Describe type of analysis and rationale for examining exclusions, including exclusion related to patient preference)***:**

2b3.3 Results *(Provide statistical results for analysis of exclusions, e.g., frequency, variability, sensitivity analyses):*

2b4. Risk Adjustment Strategy. (*For outcome measures, adjustment for differences in case mix (severity) across measured entities was appropriately tested with adequate results.*)

2b4.1 Data/Sample *(Description of the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)***:**

N/A

2b4.2 Analytic Method (*Describe methods and rationale for development and testing of risk model or risk stratification including selection of factors/variables)***:**

2b4.3 Testing Results *(Statistical risk model: Provide quantitative assessment of relative contribution of model risk factors; risk model performance metrics including cross-validation discrimination and calibration statistics, calibration curve and risk decile plot, and assessment of adequacy in the context of norms for risk models. Risk stratification: Provide quantitative assessment of relationship of risk factors to the outcome and differences in outcomes among the strata)***:**

2b4.4 **If outcome or resource use measure is not risk adjusted, provide rationale and analyses to justify lack of adjustment:** Risk adjustment is not applied for this measure at the health plan level. NCQA has determined that risk adjustment is not necessary other than the reporting of the measure is stratified by insurance coverage (commercial and Medicaid). The measure is stratified by age and product line.

2b5. Identification of Meaningful Differences in Performance. (*The performance measure scores were appropriately analyzed and discriminated meaningful differences in quality.*)

2b5.1 **Data/Sample** *(Describe the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)***:**

2b5.2 **Analytic Method** *(Describe methods and rationale to identify statistically significant and practically/meaningfully differences in performance)***:**

Comparison of means and percentiles; analysis of variance against established benchmarks; if sample size is >400, we would use an analysis of variance.

2b5.3 **Results** *(Provide measure performance results/scores, e.g., distribution by quartile, mean, median, SD, etc.; identification of statistically significant and meaningfully differences in performance)*:

2b6. Comparability of Multiple Data Sources/Methods. (*If specified for more than one data source, the various approaches result in comparable scores.*)

2b6.1 **Data/Sample** *(Describe the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)***:**

During field testing, NCQA compares performance rates based on administrative data with performance rates based on medical record review.

2b6.2 **Analytic Method** *(Describe methods and rationale for testing comparability of scores produced by the different data sources specified in the measure)***:**

Comparison of means with expert review to medical record review, which is considered the gold standard.

2b6.3 **Testing Results** *(Provide statistical results, e.g., correlation statistics, comparison of rankings; assessment of adequacy in the context of norms for the test conducted)***:**

From 2006 to 2008, the rates for women ages 16-20 varied for each reporting percentile, either remaining stable or increasing slightly, with a peak in 2007. For the 10th percentile, the rate remained relatively stable, going from 24.51% to 26.03% to 24.37%. For the 25th percentile, the rate increased slightly, going from 28.97% to to 30.94% to 29.04%. For the 50th percentile, the rate also increased slightly, going from 33.84% to to 35.97% to 34.50%. For the 75th percentile, the rate remained relatively stable, going from 39.27% to to 35.97% to 39.55%. For the 90th percentile, the rate increased, going from 43.71% to to 46.51% to 45.84%.

From 2006 to 2008, the rates for women ages 21-25 increased slightly or significantly with a peak in 2007. For the 10th

percentile, the rate increased slightly, going from 23.57% to to 25.52% to 25.32%. For the 25th percentile, the rate increased slightly, going from 28.77% to to 30.97% to 29.84%. For the 50th percentile, the rate also increased, going from 34.02% to to 37.28% to 36.50%. For the 75th percentile, the rate increased significantly, going from 39.55% to to 44.58% to 44.02%. For the 90th percentile, the rate also increased significantly, going from 47.45% to 50.45% to 51.03%.

2c. Disparities in Care: **H☐ M☐ L☐ I☐ NA☐** (*If applicable, the measure specifications allow identification of disparities.*)

2c.1 If measure is stratified for disparities, provide stratified results *(Scores by stratified categories/cohorts)*: The measure is not stratified to detect disparities.

2c.2 If disparities have been reported/identified (e.g., in 1b), but measure is not specified to detect disparities, please explain:

NCQA has participated with IOM and others in attempting to include information on disparities in measure data collection. However, at the present time, this data, at all levels (claims data, paper chart review, and electronic records), is not coded in a standard manner, and is incompletely captured. There are no consistent standards for what entity (physician, group, plan, employer) should capture and report this data. While "requiring" reporting of the data could push the field forward, it has been our position that doing so would create substantial burden with inability to use the data because of its inconsistency. At the present time, we agree with the IOM report that disparities are best considered by the use of zip code analysis which has limited applicability in most reporting situations. At the health plan level, for HEDIS health plan data collection, NCQA does have extensive data related to our use of stratification by insurance status (Medicare, Medicaid and private-commercial) and would strongly recommend this process where the data base supporting the measurement includes this information. However, we believe that the measure specifications should NOT require this since the measure is still useful where the data needed to determine disparities cannot be ascertained from the data available.

2.1-2.3 Supplemental Testing Methodology Information:

Steering Committee: Overall, was the criterion, *Scientific Acceptability of Measure Properties*, met?

(*Reliability and Validity must be rated moderate or high*) Yes☐ No☐

Provide rationale based on specific subcriteria:

If the Committee votes No, STOP

3. Feasibility

Extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

3a. Byproduct of Care Processes

For clinical measures, the required data elements are routinely generated and used during care delivery (e.g., blood pressure, lab test, diagnosis, medication order).

3a.1. **Data Elements Generated as Byproduct of Care Processes.**

Generated or collected by and used by healthcare personnel during the provision of care (e.g., blood pressure, lab value, diagnosis, depression score), Coded by someone other than person obtaining original information (e.g., DRG, ICD-9 codes on claims)

If other:

3b. Electronic Sources

The required data elements are available in electronic health records or other electronic sources. If the required data are not in electronic health records or existing electronic sources, a credible, near-term path to electronic collection is specified.

3b.1. **To what extent are the specified data elements available electronically in defined fields?** *(i.e., data elements that are needed to compute the performance measure score are in defined, computer-readable fields)*

ALL data elements are in defined fields in electronic claims

3b.2. **If ALL the data elements needed to compute the performance measure score are not from electronic sources, specify a credible, near-term path to electronic capture, OR provide a rationale for using other than electronic sources.**

3b.3. **If this is an eMeasure, provide a summary of the feasibility assessment in an attached file or make available at a measure-specific URL.**

Attachment:

3c. Data Collection Strategy

Demonstration that the data collection strategy (e.g., source, timing, frequency, sampling, patient confidentiality, costs associated with fees/licensing of proprietary measures) can be implemented (e.g., already in operational use, or testing demonstrates that it is ready to put into operational use). For eMeasures, a feasibility assessment addresses the data elements and measure logic and demonstrates the eMeasure can be implemented or feasibility concerns can be adequately addressed.

3c.1. **Describe what you have learned/modified as a result of testing and/or operational use of the measure regarding data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, time and cost of data collection, other feasibility/implementation issues.**

IF a PRO-PM, consider implications for both individuals providing PROM data (patients, service recipients, respondents) and those whose performance is being measured.

data (patients, service recipients, respondents) and those whose performance is being measured.

NCQA recognizes that, despite the clear specifications defined for HEDIS measures, data collection and calculation methods may vary, and other errors may affect the results. Thus, NCQA conducts an independent audit of all HEDIS collection and reporting processes, as well as an audit of the data which are manipulated by those processes, in order to verify that HEDIS specifications are met. NCQA has developed a precise, standardized methodology for verifying the integrity of HEDIS collection and calculation processes through a two-part program consisting of an overall information systems capabilities assessment followed by an evaluation of the MCO´s ability to comply with HEDIS specifications. NCQA-certified auditors using standard audit methodologies will help enable purchasers to make more reliable "apples-to-apples" comparisons between health plans. The HEDIS Compliance Audit addresses the following functions:

1) information practices and control procedures

2) sampling methods and procedures

3) data integrity

4) compliance with HEDIS specifications

5) analytic file production

6) reporting and documentation

In addition to the HEDIS Audit, NCQA provides a system to allow "real-time" feedback from measure users. Our Policy Clarification Support System receives thousands of inquiries each year on over 100 measures. Through this system NCQA responds immediately to questions and identifies possible errors or inconsistencies in the implementation of the measure. This system is vital to the regular re-evaluation of NCQA measures.

Input from NCQA auditing and the Policy Clarification Support System informs the annual updating of all HEDIS measures including updating value sets and clarifying the specifications. Measures are re-evaluated on a periodic basis and when there is a significant change in evidence. During re-evaluation information from NCQA auditing and Policy Clarification Support System is used to inform evaluation of the scientific soundness and feasibility of the measure.

3c.2. Describe any fees, licensing, or other requirements to use any aspect of the measure as specified *(e.g., value/code set, risk model, programming code, algorithm)*.

Broad public use and dissemination of these measures is encouraged and NCQA has agreed with NQF that noncommercial uses do not require the consent of the measure developer. Use by health care physicians in connection with their own practices is not commercial use. Commercial use of a measure requires the prior written consent of NCQA. As used herein, "commercial use" refers to any sale, license or distribution of a measure for commercial gain, or incorporation of a measure into any product or service that is sold, licensed or distributed for commercial gain, even if there is no actual charge for inclusion of the measure.

4. Usability and Use

Extent to which potential audiences (e.g., consumers, purchasers, providers, policy makers) are using or could use performance results for both accountability and performance improvement to achieve the goal of high-quality, efficient healthcare for individuals or populations.

4a. Accountability and Transparency

Performance results are used in at least one accountability application within three years after initial endorsement and are publicly reported within six years after initial endorsement (or the data on performance results are available). If not in use at the time of initial endorsement, then a credible plan for implementation within the specified timeframes is provided.

4.1. Current and Planned Use

NQF-endorsed measures are expected to be used in at least one accountability application within 3 years and publicly reported within 6 years of initial endorsement in addition to performance improvement.

Planned	Current Use (for current use provide URL)
Quality Improvement (Internal to the specific organization)	Public Reporting Health Plan Rating http://reportcard.ncqa.org/plan/external/plansearch.aspx State of Health Care Quality Report http://www.ncqa.org/tabid/836/Default.aspx Medicaid Child Core Set (16-20-year age rate) https://www.medicaid.gov/Medicaid-CHIP-Program-Information/By-Topics/Quality-of-Care/CHIPRA-Initial-Core-Set-of-Childrens-Health-Care-Quality-Measures.html Medicaid Adult Core Set: (21-24-year age rate) https://www.medicaid.gov/medicaid-chip-program-information/by-topics/quality-of-care/adult-health-care-quality-measures.html California's Value Based Pay for Performance Program: http://www.iha.org/our-work/accountability/value-based-p4p Physician Quality Reporting System: https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/PQRS/ California's Value Based Pay for Performance Program: http://www.iha.org/our-work/accountability/value-based-p4p Payment Program CMS EHR Incentive Program (Meaningful Use) https://www.healthit.gov/providers-professionals/meaningful-use-definition-objectives Physician Quality Reporting System https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/PQRS/ California's Value Based Pay for Performance Program http://www.iha.org/our-work/accountability/value-based-p4p Regulatory and Accreditation Programs Accreditation http://www.ncqa.org/tabid/123/Default.aspx HEDIS ACO http://www.ncqa.org/Programs/Accreditation/AccountableCareOrganizationACO.aspx Quality Improvement with Benchmarking (external benchmarking to multiple organizations) Quality Compass http://www.ncqa.org/tabid/177/Default.aspx

	Annual State of Health Care Quality http://www.ncqa.org/tabid/836/Default.aspx

4a.1. **For each CURRENT use, checked above, provide:**

- Name of program and sponsor
- Purpose
- Geographic area and number and percentage of accountable entities and patients included

ACCOUNTABLE CARE ORGANIZATION ACCREDITATION: This measure is used in NCQA's ACO Accreditation program, that helps health care organizations demonstrate their ability to improve quality, reduce costs and coordinate patient care. ACO standards and guidelines incorporate whole-person care coordination throughout the health care system.

CALIFORNIA'S VALUE BASED PAY FOR PERFORMANCE PROGRAM: This measure is used in the California P4P program, which is the largest non-governmental physician incentive program in the United States. Founded in 2001, it is managed by the Integrated Healthcare Association (IHA) on behalf of eight commercial HMO health plans representing 9 million insured persons. IHA reports results on approximately 35,000 physicians in 200 physician organizations.

CMS EHR INCENTIVE PROGRAM (MEANINGFUL USE): The Medicare and Medicaid Electronic Health Care Record (EHR) Incentive Programs provide incentive payments to eligible professionals, eligible hospitals, and critical access hospitals (CAHs) as they adopt, implement, upgrade or demonstrate meaningful use of certified EHR technology

NCQA HEALTH PLAN ACCREDITATION: This measure is used in scoring for accreditation of commercial and Medicaid Health Plans. In 2012, a total of 336 commercial health plans (covering 87 million lives) and 77 Medicaid health plans (covering 9.1 million lives) were accredited using this measure among others. Health plans are scored based on performance compared to benchmarks.

HEALTH PLAN RATINGS/REPORT CARDS: This measure is used to calculate health plan ratings which are reported in Consumer Reports and on the NCQA website. These ratings are based on performance on HEDIS measures among other factors. The 2015-2016 health plan ratings reviewed nearly 1,500 health plans and rated more than 1,000 private, Medicare and Medicaid health insurance plans.

MEDICAID ADULT CORE SET: These are a core set of health quality measures for Medicaid-enrolled adults. The Medicaid Adult Core Set was identified by the Centers for Medicare & Medicaid (CMS) in partnership with the Agency for HealthCare Research and Quality (AHRQ). The data collected from these measures will help CMS to better understand the quality of health care that adults enrolled in Medicaid receive nationally. Beginning in January 2014 and every three years thereafter, the Secretary is required to report to Congress on the quality of care received by adults enrolled in Medicaid. Additionally, beginning in September 2014, state data on the adult quality measures will become part of the Secretary's annual report on the quality of care for adults enrolled in Medicaid.

MEDICAID CHILD CORE SET: These are a core set of health quality measures for Medicaid-enrolled children. The Medicaid Child Core Set was identified by the Centers for Medicare & Medicaid (CMS) in partnership with the Agency for HealthCare Research and Quality (AHRQ). The data collected from these measures will help CMS to better understand the quality of health care that children enrolled in Medicaid receive nationally. On December 29, 2009, the Secretary posted for public comment in the Federal Register, an initial core set of 24 children´s health care quality measures for voluntary use by Medicaid and CHIP programs. The CHIPRA legislation provides that the Secretary shall issue updates to the Child Core Set beginning in January 2013 and annually thereafter. CMS worked with the National Quality Forum's (NQF) Measures Application Partnership (MAP) to review the Child Core Set and to identify ways to improve it. State data derived from the core measures are part of the Secretary´s Annual Report on the Quality of Care for Children in Medicaid and CHIP.

PHYSICIAN QUALITY REPORTING SYSTEM: This measure is used in the Physician Quality Reporting System (PQRS) which is a reporting program that uses a combination of incentive payments and payment adjustments to promote reporting of quality information by eligible professionals (EPs). Eligible professionals who satisfactorily report data on quality measures for covered Physician Fee

Schedule services furnished to Medicare Part B beneficiaries (including Railroad Retirement Board and Medicare Secondary Payer) receive these payment incentives and adjustments.

QUALITY COMPASS: This measure is used in Quality Compass which is an indispensable tool used for selecting a health plan, conducting competitor analysis, examining quality improvement and benchmarking plan performance. Provided in this tool is the ability to generate custom reports by selecting plans, measures, and benchmarks (averages and percentiles) for up to three trended years. Results in table and graph formats offer simple comparison of plans' performance against competitors or benchmarks.

STATE OF HEALTH CARE QUALITY REPORT: This measure is publicly reported nationally and by geographic regions in the NCQA State of Health Care annual report. This annual report published by NCQA summarizes findings on quality of care. In 2015 the report included data from 814 HMOs and 353 PPOs, representing more than 171 million patients.

4a.2. **If not currently publicly reported OR used in at least one other accountability application (e.g., payment program, certification, licensing) what are the reasons?** (*e.g., Do policies or actions of the developer/steward or accountable entities restrict access to performance results or impede implementation?*)
NA

4a.3. **If not currently publicly reported OR used in at least one other accountability application, provide a credible plan for implementation within the expected timeframes -- any accountability application within 3 years and publicly reported within 6 years of initial endorsement.** (*Credible plan includes the specific program, purpose, intended audience, and timeline for implementing the measure within the specified timeframes. A plan for accountability applications addresses mechanisms for data aggregation and reporting.*)
NA

4b. Improvement

Progress toward achieving the goal of high-quality, efficient healthcare for individuals or populations is demonstrated. If not in use for performance improvement at the time of initial endorsement, then a credible rationale describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.

4b.1. **Progress on Improvement. (Not required for initial endorsement unless available.)**
Performance results on this measure (current and over time) should be provided in 1b.2 and 1b.4. Discuss:

- **Progress (trends in performance results, number and percentage of people receiving high-quality healthcare)**
- **Geographic area and number and percentage of accountable entities and patients included**

Over the past three years, rates for this measure have remained fairly steady across both commercial and Medicaid plans . The greatest improvement in performance has been seen in commercial plans that were performing in the 10th percentile for screening women aged 21-24 for chlamydia (on average, a four percentage point improvement over time). However, performance scores demonstrate variation in the rate of women who received guideline-recommended chlamydia screening when comparing across low- and high-performing plans. For example, in 2014, there was a 23 percentage point difference between plans in the 10th percentile vs. 90th percentile among commercial plans and 30 percentage point difference among Medicaid plans for the 16-20-year rate. These gaps in performance underscore the opportunity for improvement.

4b.2. **If no improvement was demonstrated, what are the reasons? If not in use for performance improvement at the time of initial endorsement, provide a credible rationale that describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.**
While rates have been steady over the past few years, rates still indicate a gap in care. The mean rate of 45% among commercial plans and 55% among Medicaid plans shows room for improvement. Continued reporting of this measure in programs should help to focus attention on this important guideline-recommended service.

4c. Unintended Consequences

The benefits of the performance measure in facilitating progress toward achieving high-quality, efficient healthcare for individuals or populations outweigh evidence of unintended negative consequences to individuals or populations (if such evidence exists).

4c.1. **Were any unintended negative consequences to individuals or populations identified during testing; OR has evidence of unintended negative consequences to individuals or populations been reported since implementation? If so, identify the negative unintended consequences and describe how benefits outweigh them or actions taken to mitigate them.**
There were no identified unintended consequences for this measure during testing or since implementation.

5. Comparison to Related or Competing Measures

If a measure meets the above criteria and there are endorsed or new related measures (either the same measure focus or the same target population) or competing measures (both the same measure focus and the same target population), the measures are compared to address harmonization and/or selection of the best measure.

5. Relation to Other NQF-endorsed Measures
Are there related measures (conceptually, either same measure focus or target population) or competing measures (conceptually both the same measure focus and same target population)? If yes, list the NQF # and title of all related and/or competing measures.
Yes

5.1a. **List of related or competing measures (selected from NQF-endorsed measures)**
0409 : HIV/AIDS: Sexually Transmitted Diseases – Screening for Chlamydia, Gonorrhea, and Syphilis

5.1b. **If related or competing measures are not NQF endorsed please indicate measure title and steward.**

5a. Harmonization

The measure specifications are harmonized with related measures;
OR
The differences in specifications are justified

5a.1. **If this measure conceptually addresses EITHER the same measure focus OR the same target population as NQF-endorsed measure(s):**
Are the measure specifications completely harmonized?
Yes

5a.2. **If the measure specifications are not completely harmonized, identify the differences, rationale, and impact on interpretability and data collection burden.**
NQF #0409 both address chlamydia screening. However, the measures differ in the target patient populations. NQF #0409 looks for chlamydia screenings among males and females aged 13 and older with a diagnosis of HIV/AIDS. This measures focuses on women aged 16-24 with an indication of sexual activity, which aligns with the U.S. Preventive Services Task Force guideline for chlamydia screening in a general population. The measures are aligned in how they define chlamydia screening.

5b. Competing Measures

The measure is superior to competing measures (e.g., is a more valid or efficient way to measure);
OR
Multiple measures are justified.

5b.1. **If this measure conceptually addresses both the same measure focus and the same target population as NQF-endorsed measure(s):**
Describe why this measure is superior to competing measures (e.g., a more valid or efficient way to measure quality); OR provide a rationale for the additive value of endorsing an additional measure. (Provide analyses when possible.)
NA

S.12. **Stratification Details/Variables** *(All information required to stratify the measure results including the stratification variables, definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format with at S.2b)*
N/A

S.13. **Risk Adjustment Type** (Select type. Provide specifications for risk stratification in S.12 and for statistical model in S.14-15)
Statistical risk model
If other:

S.14. **Identify the statistical risk model method and variables** *(Name the statistical method - e.g., logistic regression and list all the*

risk factor variables. Note - risk model development and testing should be addressed with measure testing under Scientific Acceptability)
Logistic regression with shrinkage estimate - see S. 15a

S.15. **Detailed risk model specifications** *(must be in attached data dictionary/code list Excel or csv file. Also indicate if available at measure-specific URL identified in S.1.)*
Note: Risk model details (including coefficients, equations, codes with descriptors, definitions), should be provided on a separate worksheet in the suggested format in the Excel or csv file with data dictionary/code lists at S.2b.
Provided in response box S.15a

S.15a. **Detailed risk model specifications** (*if not provided in excel or csv file at S.2b*)
Logistic regression is used to model the infection measure on the following covariates: gestational age and its squared term; small for gestational age (Yes/No); multiple gestation (Yes/No); APGAR score at 1 minute (0-10); infant sex (Female, Male); vaginal delivery (Yes/No); major congenital malformation (Yes/No); and birth location (Inborn, Outborn). An estimate of the systematic variation associated with the hospital standardized morbidity ratios (SMRs) shrinks center SMR values and their confidence limits based on the number of infants reported (see, e.g., Martuzzi M and Hills M. Estimating the degree of heterogeneity between event rates using likelihood. American Journal of Epidemiology 1995; 141(4): 369-374; Simpson J et al. Analysing differences in clinical outcomes between hospitals. Quality and Safety in Health Care 2003; (12): 257-262).

Appendix

A.1 Supplemental materials may be provided in an appendix. All supplemental materials (such as data collection instrument or methodology reports) should be organized in one file with a table of contents or bookmarks. If material pertains to a specific submission form number, that should be indicated. Requested information should be provided in the submission form and required attachments. There is no guarantee that supplemental materials will be reviewed.

No appendix **Attachment:**

Contact Information

Co.1 **Measure Steward (Intellectual Property Owner):** National Committee for Quality Assurance
Co.2 **Point of Contact:** Bob, Rehm, nqf@ncqa.org, 202-955-1728-
Co.3 **Measure Developer if different from Measure Steward:** National Committee for Quality Assurance
Co.4 **Point of Contact:** Bob, Rehm, nqf@ncqa.org, 202-955-3599-

Additional Information

Ad.1 **Workgroup/Expert Panel involved in measure development**
Provide a list of sponsoring organizations and workgroup/panel members' names and organizations. Describe the members' role in measure development.
Committee on Performance Measurement (CPM)

Bruce Bagley, MD, American Academy of Family Physicians

Andrew Baskin, MD, Aetna

Patrick Conway, MD, MSC, Center for Medicare & Medicaid Services

Jonathan D. Darer, MD, Geisinger Health System

Helen Darling, National Business Group on Health

Rebekah Gee, MD, MPH, FACOG, LSU School of Medicine and Public Health

Foster Gesten, MD, NYSDOH Office of Managed Care

David Grossman, MD, MPH, Group Health Physicians

Christine Hunter, MD (Co-Chair), US Office of Personnel Management

Jeffrey Kelman, MMSc, MD, Centers for Medicare & Medicaid Services

Bernadette Loftus, MD, The Permanente Medical Group

J. Brent Pawlecki, MD MMM, The Goodyear Tire & Rubber Company

Susan Reinhard, RN, PhD, AARP

Eric C. Schneider, MD, MSc (Co-Chair), RAND Corporation

Marcus Thygeson, MD, MPH Blue Shield of Califorina

HEDIS Expert Coding Panel

Glen Braden, MBA, CHCA, Attest Health Care Advisors, LLC

Denene Harper, RHIA, American Hospital Association

DeHandro Hayden, BS, American Medical Association

Patience Hoag, RHIT, CPHQ, CHCA, CCS, CCS-P, Health Services Advisory Group

Nelly Leon-Chisen, RHIA, American Hospital Association

Tammy Marshall, LVN, Aetna

Alec McLure, RHIA, CCS-P, Verisk Health

Michele Mouradian, RN, BSN, McKesson Health Solutions

Craig Thacker, RN, CIGNA HealthCare

Mary Jane F. Toomey, RN CPC, Aetna Better Health

NCINQ Measurement Advisory Panel

Mary Applegate, MD, Ohio Department of Job and Family Services

Katie Brookler, Colorado Department of Health Care Policy and Financing

Cathy Caldwell, MPH, Alabama Department of Public Health

Ted Ganiats, MD, University of California, San Diego

Darcy Gruttadaro, JD, National Allegiance on Mental Illness

Jennifer Havens, MD, NYU School of Medicine

Virginia Moyer, MD, MPH, FAAP, Baylor College of Medicine, USPSTF

Edward Schor, MD, Lucile Packard Foundation for Children's Health

Xavier Sevilla, MD, FAAP, Whole Child Pediatrics

Gwen Smith, Illinois Department of Healthcare and Family Services/Health Management Associates

Janet (Jessie) Sullivan, MD, Hudson Health Plan

Kalahn Taylor-Clark, PhD, MPH, George Mason University

Craig Thiele, MD, CareSource

Jeb Weisman, PhD, Children's Health Fund

Charles Wibbelsman, MD, Kaiser Permanente Medical Group, Inc.

NCINQ Clinician Advisory Panel

Elizabeth Alderman, MD, FAAP, Albert Einstein College of Medicine

Sarah Brewington, MD, Sandhills Pediatrics Inc

Gale Burstein, MD, MPH, FAAP, FSAHM, Women and Children's Hospital of Buffalo, NY

Barry Bzostek, MD, FAAP, Women and Children's Hospital of Buffalo, NY

Danielle Casher, MD, FAAP, St. Christoper's Hospital for Children

Edward Curry, MD, FAAP, Emergency Department, St. Christopher's Hospital for Children, PA

Eve Kimball, MD, FAAP, Southern California Permanente Medical Group

Paul Melinkovich, MD, FAAP, Kaiser Permanente

Jackie Nelson, MD, FAAP, Lander Regional H

Ellen Squire, MD, FAAP, HaysMed Pediatric Center

Measure Developer/Steward Updates and Ongoing Maintenance

Ad.2 **Year the measure was first released:** 1999

Ad.3 **Month and Year of most recent revision:** 2010

Ad.4 **What is your frequency for review/update of this measure?** Approximately every 3 years, sooner if the clinical guidelines have changed significantly.

Ad.5 **When is the next scheduled review/update for this measure?** 12, 2017

Ad.7 **Disclaimers:** These performance measures are not clinical guidelines and do not establish a standard of medical care, and have not been tested for all potential applications.

THE MEASURES AND SPECIFICATIONS ARE PROVIDED "AS IS" WITHOUT WARRANTY OF ANY KIND.

Ad.8 **Additional Information/Comments:** NCQA Notice of Use. Broad public use and dissemination of these measures is encouraged and NCQA has agreed with NQF that noncommercial uses do not require the consent of the measure developer. Use by health care physicians in connection with their own practices is not commercial use. Commercial use of a measure requires the prior written

MEASURE WORKSHEET

This document summarizes the evaluation of the measure as it progresses through NQF's Consensus Development Process (CDP). The information submitted by measure developers/stewards is included after the Brief Measure Information, Preliminary Analysis, and Pre-meeting Public and Member Comments sections.

To navigate the links in the worksheet: Ctrl + click link to go to the link; ALT + LEFT ARROW to return

Brief Measure Information

NQF #: 0304
Measure Title: Late sepsis or meningitis in Very Low Birth Weight (VLBW) neonates (risk-adjusted)
Measure Steward: Vermont Oxford Network
Brief Description of Measure: Standardized morbidity ratio and observed minus expected measure for nosocomial bacterial infection after day 3 of life in very low birth weight infants
Developer Rationale: A bundle of improvement practices has been shown to dramatically reduce the frequency of hospital acquired infections for very low birth weight infants.

Numerator Statement: Eligible infants with one or more of the following criteria:
Criterion 1:
Bacterial Pathogen. A bacterial pathogen is recovered from a blood and/or cerebral spinal fluid culture obtained after Day 3 of life.
OR
Criterion 2:
Coagulase Negative Staphylococcus. The infant has all 3 of the following:
1. Coagulase negative staphylococcus is recovered from a blood culture obtained from either a central line, or peripheral blood sample and/or is recovered from cerebrospinal fluid obtained by lumbar puncture, ventricular tap or ventricular drain.
2. One or more signs of generalized infection (such as apnea, temperature instability, feeding intolerance, worsening respiratory distress or hemodynamic instability).
3. Treatment with 5 or more days of intravenous antibiotics after the above cultures were obtained. If the infant died, was discharged, or transferred prior to the completion of 5 days of intravenous antibiotics, this condition would still be met if the intention were to treat for 5 or more days.
Denominator Statement: Eligible infants who are in the reporting hospital after day 3 of life.
Denominator Exclusions: Infants who do not meet eligibility criteria for birth weight, gestational age or hospital admission, or if the infant is discharged home, is transferred or dies prior to day 3 of life.

Measure Type: Outcome
Data Source: Electronic Clinical Data: Registry
Level of Analysis: Facility

IF Endorsement Maintenance – Original Endorsement Date: Nov 15, 2007 **Most Recent Endorsement Date:** Apr 02, 2012

Maintenance of Endorsement -- Preliminary Analysis

To maintain NQF endorsement endorsed measures are evaluated periodically to ensure that the measures still meets the NQF endorsement criteria ("maintenance"). The emphasis for maintaining endorsement is focused on how effective the measure is for promoting improvements in quality. Endorsed measures should have some experience from the field to inform the evaluation. The emphasis for maintaining endorsement is noted for each criterion.

Criteria 1: Importance to Measure and Report

1a. Evidence

Maintenance measures – less emphasis on evidence unless there is new information or change in evidence since the prior evaluation.

1a. Evidence. The evidence requirements for a health outcomes measure include providing rationale that supports the relationship of the health outcome to processes or structures of care. The guidance for evaluating the clinical evidence asks if the relationship between the measured health outcome and at least one clinical action is identified and supported by the stated rationale.

Evidence Summary of prior review in 2012

The developer provides 4 observational or quasi-experimental studies that support their statement "quality improvement interventions targeting appropriate practices such as improving hand hygiene and central line management can reduce hospital acquired infections in the neonatal intensive care unit (NICU)".

Changes to evidence from last review

- ☐ **The developer attests that there have been no changes in the evidence since the measure was last evaluated.**
- ☒ **The developer provided updated evidence for this measure:**

Updates: The developer provided update: 11 observational or quasi-experimental studies and one clinical guideline: Strategies for prevention of health care-associated infections in the NICU. The guideline lists specific recommendations for each of the following strategies: hand hygiene; prevention of central line-associated bloodstream infections; prevention of health care-associated pneumonia; skin care; use of human milk feedings; and antibiotic stewardship.

Question for the Committee:

- *Although the guidelines have been updated/new studies have been provided, the underlying evidence presented appears to be the directionally the same/stronger since the last NQF endorsement review. Does the Committee agree and so there is no need for repeat discussion and vote on Evidence?*

Guidance from algorithm #1 Evaluating Clinical Evidence: Health outcome or PRO (Box 1) → Relationship between the measured health outcome/PRO (Box 2) → Rate as pass

Preliminary rating for evidence: ☒ **Pass** ☐ **No Pass**

1b. Gap in Care/Opportunity for Improvement and 1b. Disparities

Maintenance measures – increased emphasis on gap and variation

1b. Performance Gap. The performance gap requirements include demonstrating quality problems and opportunity for improvement.

The developer presents the following data from Vermont-Oxford Network 2006-2014 at the facility level.

Year	N Hospitals	Mean	Minimum	Maximum
2006	672	0.192	0.000	1.000
2007	724	0.18	0.000	1.000
2008	799	0.172	0.000	1.000
2009	839	0.157	0.000	1.000
2010	789	0.139	0.000	1.000
2011	793	0.132	0.000	1.000
2012	910	0.12	0.000	1.000
2013	921	0.112	0.000	0.677
2014	938	0.108	0.000	0.813

Disparities

The developer presents measure results stratified by race/ethnicity using information from the Vermont Oxford Network: 2006 - 672 hospitals
2014 - 938 hospitals

Year	White non-Hispanic	Black non-Hispanic	Hispanic	Asian
2006	0.195	0.234	0.215	0.176
2007	0.186	0.225	0.208	0.158
2008	0.178	0.201	0.192	0.143
2009	0.164	0.184	0.175	0.138
2010	0.147	0.160	0.160	0.125
2011	0.124	0.145	0.146	0.124
2012	0.123	0.139	0.127	0.104
2013	0.118	0.122	0.126	0.093
2014	0.116	0.126	0.125	0.109

The developer also states "Rates of hospital acquired bacterial infection varied by race/ethnicity of the mother, ranging from 11.4% for infants with black mothers to 8.9% for infants with Asian mothers."

Questions for the Committee:

- *Is there a gap in care that warrants a national performance measure?*
- *Does this measure provide information about disparities of care?*

Preliminary rating for opportunity for improvement: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 1: Importance to Measure and Report (including 1a, 1b, 1c)

1a. Evidence to Support Measure Focus

Comments:

The measure relates to an outcome of sepsis or meningitis in VLBW infants. There is a direct relationship between this outcome and processes fro prevention of blood stream infectiosn, eg-handwashing, care and maintenance of central line and other processes of care.

This is a measure of a health outcome. The measure steward provided a list of 15 studies and guidelines that link adherence to processes of care and improved outcomes.

1b. Performance Gap

Comments:

There is a substantial variation in the outcome measure within the VON member organizations. About 10% institutions report a below expected sepsis rate and a similar percentage report an above average rate.

Performance on this measure has improved over the last 8 years, including reductions in disparities. There still remains the opportunity for improvement on this measure.

1c. High Priority (previously referred to as High Impact)

Comments:

Yes.

Not applicable.

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability

2a1. Reliability Specifications
Maintenance measures – no change in emphasis – specifications should be evaluated the same as with new measures

2a1. Specifications requires the measure, as specified, to produce consistent (reliable) and credible (valid) results about the quality of care when implemented.

Data source: Electronic Clinical Registry Data from the Vermont Oxford Network Database.

Specifications:

- The level of analysis is at the facility and the care setting is hospital/acute care facility.
- The numerator is Infants whose birth weight is between 401 and 1500 grams or whose gestational age is between 22 weeks 0 days and 29 weeks 6 days with bacterial infection or signs of generalized infection.
- The denominator includes infants whose birth weights are between 401 and 1500 grams or whose gestational ages are between 22 weeks 0 days and 29 weeks 6 days who are in the reporting hospital after day 3 of life.
- The denominator exclusions include infants who do not meet eligibility criteria for birth weight, gestational age, hospital admission, if the infant is discharged home, is transferred, or dies prior to day 3 of life.
- An attached spreadsheet contains numerous ICD-9 and ICD-10 codes for newborn gestational age, newborn birth weight, and sepsis, however, VON does not use ICD-9 or ICD-10 codes in its definitions and that the supplied codes have not been reviewed or vetted.
- The developer reports there have been no changes to the specifications.
- The measure is risk-adjusted using logistic regression with shrinkage estimate.
- A calculation algorithm describes the process of calculating the measure.

Questions for the Committee:

- *Are all the data elements clearly defined? Are all appropriate codes included?*
- *Is the logic or calculation algorithm clear?*
- *Is it likely this measure can be consistently implemented?*

2a2. Reliability Testing Testing attachment
Maintenance measures – less emphasis if no new testing data provided

2a2. Reliability testing demonstrates if the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise enough to distinguish differences in performance across providers.

For maintenance measures, summarize the reliability testing from the prior review:
Specific reliability testing beyond the performance of the statistical model was not discussed. Logistic regression models were used to evaluate the measure performance using the area under the receiver operating characteristic curve and monitored over time.

Describe any updates to testing see below

SUMMARY OF TESTING
Reliability testing level ☒ **Measure score** ☐ **Data element** ☐ **Both**
Reliability testing performed with the data source and level of analysis indicated for this measure ☐ **Yes** ☒ **No**

Method of reliability testing:

- The developer reports on a split-half analysis to measure the internal consistency of the measure. They selected hospitals with at least 10 infants, ran 100 samples of patients within each hospital randomly split to two groups, computed hospital rates, and calculated the correlations.

Results of reliability testing:

- Aggregate reliability for 2006 – 2014 is provided. The reliability was 0.63 in 2014. Reliability estimates range from 0 (no correlation) to 1 (perfect correlation). Split-half is an appropriate test of reliability.
- The developers conclude that "The correlation coefficients were lower than we expected. It suggests that the definition may not be applied in the same manner across all infants at all hospitals. We did notice that the

coefficients increased as the number of infants at the hospitals increased."

Guidance from the Reliability Algorithm

Precise specifications (Box 1) → empirical testing (box2) → testing of measure score (Box 4) → appropriate method (Box 5) → moderate confidence scores are reliable (Box 6b) →moderate

Questions for the Committee:

- *Is the test sample adequate to generalize for widespread implementation?*
- *Do the results demonstrate sufficient reliability so that differences in performance can be identified?*

Preliminary rating for reliability: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

2b. Validity
Maintenance measures – less emphasis if no new testing data provided

2b1. Validity: Specifications

2b1. Validity Specifications. This section should determine if the measure specifications are consistent with the evidence.

Specifications consistent with evidence in 1a. ☒ **Yes** ☐ **Somewhat** ☐ **No**

Question for the Committee:

- *Are the specifications consistent with the evidence?*

2b2. Validity testing

2b2. Validity Testing should demonstrate the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality.

For maintenance measures, summarize the validity testing from the prior review:

Assessment of adequate face validity. The prior Committee did not voice any concerns.

Describe any updates to validity testing – none.

SUMMARY OF TESTING

Validity testing level ☒ **Measure score** ☐ **Data element testing against a gold standard** ☐ **Both**

Method of validity testing of the measure score:

- ☒ **Face validity only**
- ☐ **Empirical validity testing of the measure score**

Validity testing method:

Face validity was assessed using a panel of clinical experts in neonatal infection. The measure is reviewed annually by the Vermont Oxford Network Database Advisory Committee, consisting of national and international experts in the neonatal community.

Validity testing results:

This measure was deemed valid by the developer's expert panel.

2b3-2b7. Threats to Validity

2b3. Exclusions:

- The denominator exclusions include infants who do not meet eligibility following criteria:
  - Any infant who meets neither of the following conditions is excluded:
    - Birth weight between 401 and 1500 grams
    - Gestational age between 22 and 29 weeks.
  - Outborn infants who are admitted to the reporting hospital more than 28 days after birth are excluded.
  - Outborn infants who have been home prior to admission to the reporting hospital are excluded.
  - Infants discharged home on or before day 3 of life are excluded.
  - Infants who die on or before day 3 of life are excluded.
  - Infants who transfer to another hospital on or before day 3 of life and who are not readmitted to the reporting hospital.
  - Infants who transfer more than once prior to day 3 of life.

- The developer provides an analysis on the frequency of exclusions:

Year	Number of infants	Number of infants excluded	% excluded
2006	41657	1921	4.6
2007	49866	2196	4.4
2008	52799	2151	4.1
2009	54042	2275	4.2
2010	53500	2052	3.8
2011	55170	2180	4.0
2012	55638	2110	3.8
2013	56168	2152	3.8
2014	56730	2186	3.9

Questions for the Committee:

- *Are the exclusions consistent with the evidence?*
- *Are any patients or patient groups inappropriately excluded from the measure?*
- *Are the exclusions/exceptions of sufficient frequency and variation across providers to be needed (and outweigh the data collection burden)?*

2b4. Risk adjustment: **Risk-adjustment method** ☐ **None** ☒ **Statistical model** ☐ **Stratification**

- Logistic regression models were used to evaluate the measure performance using the area under the receiver operating characteristic curve and monitored over time.

Conceptual rationale for SDS factors included? ☐ **Yes** ☒ **No**

SDS factors included in risk model? ☐ **Yes** ☒ **No**

Risk adjustment summary

The developer used Hosmer-Lemeshow goodness of fit statistics to evaluate their statistical risk model. The risk model includes eight risk factors, which are characteristics of the infants at birth.

Statistical Risk Model Calibration Statistics

Birth Year	Fit Chi-Square
2006	26.0
2007	41.9
2008	18.0
2009	37.8
2010	30.6
2011	35.4
2012*	14.8

2013	17.9
2014	13.5

Questions for the Committee:

- *Is an appropriate risk-adjustment strategy included in the measure?*
- *Are the candidate and final variables included in the risk adjustment model adequately described for the measure to be implemented?*
- *Are all of the risk adjustment variables present at the start of care? If not, describe the rationale provided.*
- *Do you agree with the developer's rationale that there is no conceptual basis for adjusting this measure for SDS factors?*
- *Do you agree with the developer's decision, based on their analysis, to not include SDS factors in their risk-adjustment model?*

2b5. Meaningful difference *(can statistically significant and clinically/practically meaningful differences in performance measure scores can be identified)*:

Here are the classifications used the developer:

- 95% upper bound for the standardized morbidity ratio (SMR - observed value / expected value) is less than one, the hospital performance is classified as "better than expected"
- 95% lower bound for the SMR is greater than one, the hospital performance is classified as "worse than expected"
- 95% lower and upper bounds for the SMR includes one, performance is classified as "as expected."

In 2014 of the 932 eligible hospitals:

#Hospitals (2014	Performance
770	performed as expected (i.e., the confidence interval contained 1)
92	performed worse than expected (i.e., the lower bound was greater than 1)
70	performed better than expected (i.e., the upper bound was less than 1).

The developer expects the majority of hospitals to perform 'as expected' but does note that there will be outliers.

Question for the Committee:

- *Does this measure identify meaningful differences about quality?*

2b6. Comparability of data sources/methods: single data source – VON registry

2b7. Missing Data

- The developer notes that over the course of nine years 0.08% or 378 infants had missing data.

Guidance from algorithm #3 Evaluating Validity: Specifications consistent with evidence (Box 1) → Assessed all potential threats to validity (Box 2) → Face validity systematically assessed→ moderate (in the absence of empirical validity testing, the highest rating possible is moderate.)

Question for the Committee:

- *Since the developer does not provide any new validity testing aside from updating the risk model discrimination, does the Committee accept the prior evaluation on this criterion without need for further discussion and voting?*

Preliminary rating for validity: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 2: Scientific Acceptability of Measure Properties (including all 2a, 2b, and 2d)

2a1. & 2b1. Specifications

Comments:

The data definitions are provided unambiguously. However, there is confusion about the numerator. It is stated in the application that VON collects data on infants with BW between 501 g and 1500 g, whereas the measure applies to infants with BW between 401 g and 1500 g.

Concerns that portions of Criterion 2 could be a bit vague (e.g. signs of generalized infection)

Not inconsistent.

2a2. Reliability Testing

Comments:

**Moderate. The reliability was 0.63 (with a range from 0 to 1). The submission states that the correlation coefficients were lower than expected. The coefficients increased as number of infants in the hospital increased. **

Steward ran split-half analysis; reliability was 0.63. Based on algorithm, qualifies for 'moderate' rating. Developer acknowledges that the definitions may not be applied in the same manner across all infants at all hospitals.

2b2. Validity Testing

Comments:

Face validity assessed by a panel of experts in Neonatal Perinatal medicine.

Assessed using face validity, using a panel of clinical experts in neonatal infection. Measure is reviewed annually by VON's advisory committee. No data on other types of validity testing.

2b3. Exclusions Analysis

2b4. Risk Adjustment/Stratification for Outcome or Resource Use Measures

2b5. Identification of Statistically Significant & Meaningful Differences In Performance

2b6. Comparability of Performance Scores When More Than One Set of Specifications

2b7. Missing Data Analysis and Minimizing Bias

Comments:

Exclusions are appropriate. SDS were not controlled for or included. Not including SDS appears appropriate as SDS should not influence this outcome. Missing data are not a relevant threat to validity of this measure.

No. There is very little missing data (0.08% missing data rate over 9 year period).

Criterion 3. Feasibility

Maintenance measures – no change in emphasis – implementation issues may be more prominent

3. Feasibility is the extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

- This measure is based on clinical registry data and all data elements are available in electronic sources.
- All data is in electronic format in a clinical registry.
- There are no fees to use the measure; however, members of the Vermont Oxford Network pay an annual membership fee.

Questions for the Committee:

- *Are the required data elements routinely generated and used during care delivery?*
- *Is this registry-based measure available for use by everyone?*

Preliminary rating for feasibility: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 3: Feasibility

3a. Byproduct of Care Processes
3b. Electronic Sources
3c. Data Collection Strategy

Comments:

The data are collected as a part of the VON registry. Only member institutions will be able to have access to this data and its use.
Collected through a registry. All data elements are available through electronic sources.

Criterion 4: Usability and Use
Maintenance measures – increased emphasis – much greater focus on measure use and usefulness, including both impact /improvement and unintended consequences

4. Usability and Use evaluate the extent to which audiences (e.g., consumers, purchasers, providers, policymakers) use or could use performance results for both accountability and performance improvement activities.

Current uses of the measure
Publicly reported? ☐ **Yes** ☒ **No**

Current use in an accountability program? ☐ **Yes** ☒ **No**
OR
Planned use in an accountability program? ☐ **Yes** ☒ **No**

Accountability program details

- This measure has been NQF endorsed since 2007. NQF criteria for usability and use is looking for "performance results are used in at least 1 accountability application within 3 years after initial endorsement and are publicly reported within 6 years after initial endorsement (or the data on performance results are available).
- The developer states "Vermont Oxford Network is committed to working with accrediting bodies that are developing publicly-reported quality measures for the neonatal population."

Improvement results

- Data over the years 2007 to 2013 78% of Vermont Oxford Network member hospitals reported decreases in rates infection.

Unexpected findings (positive or negative) during implementation

- To mitigate unintended consequences, the developer explains that hospitals receive a manual of operations annually that contains definitions and clearly operationalized criteria for the measure. Comprehensive business rules verify records for consistency, completeness and accuracy. Centers employ a definitive process to assure that the measure is not reported until data are complete and correct. Hospital contacts must verify that data for all eligible infants are submitted prior to finalization

Potential harms: none reported

Feedback:

***Questions for the Committee**:*

- *Despite being endorsed since 2007 this measure is neither publicly reported nor used in an accountability*

program. Does the Committee think that either is likely in the near future?

- *How can the performance results be used to further the goal of high-quality, efficient healthcare?*
- *Do the benefits of the measure outweigh any potential unintended consequences?*

Preliminary rating for usability and use: ☐ **High** ☐ **Moderate** ☒ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 4: Usability and Use

4a. Accountability and Transparency

4b. Improvement

4c. Unintended Consequences

Comments:

This measure is not being publically reported. There are no publicly reported outcome measures for VLBW infants.
**The measure is not publicly reported or used in accountability program, despite being endorsed since 2007.
VON states they are committed to working with accrediting bodies that are developing publicly-reported quality measures for the neonatal population.**

Criterion 5: Related and Competing Measures

Related or competing measures

- 0478: Neonatal Blood Stream Infection Rate (NQI 03)
- 1731: PC-04 Health Care-Associated Bloodstream Infections in Newborns

Harmonization

- Currently there are no efforts to harmonize these measures. The developer explains that the measures have different target populations, different item definitions, and also different risk adjustment methodology.

Pre-meeting public and member comments

-

1. IMPACT, OPPORTUITY, EVIDENCE - IMPORTANCE TO MEASURE AND REPORT

Importance to Measure and Report is a threshold criterion that must be met in order to recommend a measure for endorsement. All three subcriteria must be met to pass this criterion. See guidance on evidence.

Measures must be judged to be important to measure and report in order to be evaluated against the remaining criteria. (evaluation criteria)

1c.1 Structure-Process-Outcome Relationship *(Briefly state the measure focus, e.g., health outcome, intermediate clinical outcome, process, structure; then identify the appropriate links, e.g., structure-process-health outcome; process- health outcome; intermediate clinical outcome-health outcome)*:
Health outcome: hospital acquired bacterial infection
Process: specific practices related to hand hygiene, line insertion, care and removal
Structure: unit culture
Links: unit culture impacts adherence to infection prevention practices which influence rate of infection

1c.2-3 Type of Evidence *(Check all that apply)*:
Clinical Practice Guideline
Selected individual studies (rather than entire body of evidence)

1c.4 Directness of Evidence to the Specified Measure *(State the central topic, population, and outcomes addressed in the body of evidence and identify any differences from the measure focus and measure target population)*:
There is strong evidence that quality improvement interventions targeting appropriate practices such as improving hand hygiene and central line management can reduce hospital acquired infections in the neonatal intensive care unit (NICU).

1c.5 Quantity of Studies in the Body of Evidence *(Total number of studies, not articles)*: numerous

1c.6 Quality of Body of Evidence (*Summarize the certainty or confidence in the estimates of benefits and harms to patients across studies in the body of evidence resulting from study factors. Please address: a) study design/flaws; b) directness/indirectness of the evidence to this measure (e.g., interventions, comparisons, outcomes assessed, population included in the evidence); and c) imprecision/wide confidence intervals due to few patients or events):*
The majority of the available studies from NICUs are observational or quasi-experimental.

1c.7 Consistency of Results across Studies *(Summarize the consistency of the magnitude and direction of the effect):* Consistent in magnitude and direction across NICU studies and when compared to similar studies in adult and pediatric intensive care.

1c.8 Net Benefit *(Provide estimates of effect for benefit/outcome; identify harms addressed and estimates of effect; and net benefit - benefit over harms)*: Marked reductions in hospital acquired bacterial infections in the NICU can lead to better outcomes, shorter hospital stay, and lower costs while in the hospital as well as better developmental outcomes at follow-up.

1c.9 Grading of Strength/Quality of the Body of Evidence. Has the body of evidence been graded? No

1c.10 If body of evidence graded, identify the entity that graded the evidence including balance of representation and any disclosures regarding bias: N/A

1c.11 System Used for Grading the Body of Evidence: Other

1c.12 If other, identify and describe the grading scale with definitions: N/A

1c.13 Grade Assigned to the Body of Evidence: N/A

1c.14 Summary of Controversy/Contradictory Evidence:
Given that the evidence is predominantly observational, there is the possibility that the magnitude of effect of quality improvement interventions on hospital acquired infection could be confounded by the non-randomized nature of the studies.

1c.15 Citations for Evidence other than Guidelines*(Guidelines addressed below)*:

Updated using articles published since last NQF submission

Bizzarro Matthew J. B. Sabo, M. Noonan, M.P. Bonfiglio, V. Northrup, K. Diefenbach; Central Venous Catheter Initiative Committee. "A quality improvement initiative to reduce central line-associated bloodstream infections in a neonatal intensive care unit." *Infection Control and Hospital Epidemiology* 31, no. 3 (2010): 241-248.

Schulman, Joseph, Rachel Stricof, Timothy P. Stevens, Michael Horgan, Kathleen Gase, Ian R. Holzman, Robert I. Koppel, Suhas Nafday, Kathleen Gibbs, Robert Angert, Aryeh Simmonds, Susan A. Furdon, Lisa Saiman, and the New York State Regional Perinatal Care Centers. "Statewide NICU central-line-associated bloodstream infection rates decline after bundles and checklists." *Pediatrics* 127, no. 3 (2011): 436-444.

Wirtschafter, David D., Richard J. Powers, Janet S. Pettit, Henry C. Lee, W. John Boscardin, Mohammad Ahmad Subeh, and Jeffrey B. Gould. "Nosocomial infection reduction in VLBW infants with a statewide quality-improvement model." *Pediatrics* 127, no. 3 (2011): 419-426.

Kaplan, Heather C., Carole Lannon, Michele C. Walsh, Edward F. Donovan, and for the Ohio Perinatal Quality Collaborative. Pediatrics. "Ohio statewide quality-improvement collaborative to reduce late-onset sepsis in preterm infants." Pediatrics 127, no. 3 (2011): 427-435.

Holzmann-Pazgal, G., A. Kubanda, K. Davis, A. M. Khan, K. Brumley, and S. E. Denson. "Utilizing a line maintenance team to reduce central-line-associated bloodstream infections in a neonatal intensive care unit." *Journal of Perinatology* 32, no. 4 (2012): 281-286.

Fisher, David, Keith M. Cochran, Lloyd P. Provost, Jacquelyn Patterson, Tara Bristol, Karen Metzguer, Brian Smith, Daniela Testoni, and Martin J. McCaffrey. "Reducing Central Line–Associated Bloodstream Infections in North Carolina NICUs." *Pediatrics* 132, no. 6 (2013): e1664-e1671.

Milstone, Aaron M., Nicholas G. Reich, Sonali Advani, Guoshu Yuan, Kristina Bryant, Susan E. Coffin, W. Charles Huskins et al. "Catheter dwell time and CLABSIs in neonates with PICCs: a multicenter cohort study." *Pediatrics* 132, no. 6 (2013): e1609-e1615.

Patrick, Stephen W., Matthew M. Davis, Aileen B. Sedman, Jennifer A. Meddings, Sue Hieber, Grace M. Lee, Terri L. Stillwell, Carol E. Chenoweth, Claudia Espinosa, and Robert E. Schumacher. "Accuracy of hospital administrative data in reporting central line–associated bloodstream infections in newborns." *Pediatrics* 131, no. 1 (2013): S75-S80.

Ceballos, Kirtley, Kari Waterman, Teresa Hulett, and Mary Beth Flynn Makic. "Nurse-driven quality improvement interventions to reduce hospital-acquired infection in the NICU." *Advances in Neonatal Care* 13, no. 3 (2013): 154-163.

Quach, Caroline, Aaron M. Milstone, Chantal Perpête, Mario Bonenfant, Dorothy L. Moore, and Therese Perreault. "Chlorhexidine bathing in a tertiary care neonatal intensive care unit: impact on central line–associated bloodstream infections." *Infection Control* 35, no. 02 (2014): 158-163.

Zachariah Philip, E. Yoko Furuya, Jeffrey Edwards, Andrew Dick, Hangsheng Liu, Carolyn Herzig, Monika Pogorzelska-Maziarz, Patricia W. Stone, Lisa Saiman. "Compliance with prevention practices and their association with central line-associated blood stream infections in neonatal intensive care units." *American Journal of Infection Control*. 42, no. 8 (2014): 847-851.

Shepherd, Edward G., Tami J. Kelly, Jodi A. Vinsel, Dennis J. Cunningham, Erin Keels, Wendi Beauseau, and Richard E. McClead. "Significant reduction of central-line associated bloodstream infections in a network of diverse neonatal nurseries." *The Journal of Pediatrics* 167, no. 1 (2015): 41-46.

Greenberg, Rachel G., Keith M. Cochran, P. Brian Smith, Barbara S. Edson, Joseph Schulman, Henry C. Lee, Balaji Govindaswami, Alfonso Pantoja, Doug Hardy, John Curran, Della Lin, Sheree Kuo, Akihiko Noguchi, Patricia Ittmann, Scott Duncan, Munish Gupta, Alan Picarillo, Padmani Karna, Morris Cohen, Michael Giuliano, Sheri Carroll, Brandi Page, Judith Guzman-Cottrill, M. Whit Walker, Jeff Garland, Janice K. Ancona, Dan L. Ellsbury, Matthew M. Laughon, Martin J. McCaffrey. "Effect of catheter dwell time on risk of central line-associated bloodstream infection in infants." *Pediatrics* 136, no. 6 (2015): 1080-1086.

Piazza, Anthony J., Beverly Brozanski, Lloyd Provost, Theresa R. Grover, John Chuo, Joan R. Smith, Teresa Mingrone, Susan Moran, Lorna Morelli, Isabella Zaniletti, Eugenia K. Pallotto. "SLUG Bug: Quality improvement with orchestrated testing leads to NICU CLABSI reduction." *Pediatrics* 137, no 1 (2016): 1-12.

Popoola, Victor O., Elizabeth Colantuoni, Nuntra Suwantarat, Rebecca Pierce, Karen C. Carroll, Susan W. Aucott, Aaron M. Milstone. "Active surveillance cultures and decolonization to reduce Staphylococcus aureus infections in the neonatal intensive care unit." *Infection Control and Hospital Epidemiology* (2016), epub ahead of print.

1c.16 Quote verbatim, the specific guideline recommendation *(Including guideline # and/or page #)*:
The guideline lists specific recommendations for each of the following strategies: hand hygiene; prevention of central line-associated bloodstream infections; prevention of health care-associated pneumonia; skin care; use of human milk feedings; and antibiotic stewardship.

1c.17 Clinical Practice Guideline Citation:
Polin, Richard A. Susan Denson, Michael T. Brady, The Committee on Fetus and Newborn and Committee on Infectious Diseases. "Strategies for prevention of health care-associated infections in the NICU." *Pediatrics* 129, no. 4 (2012): e1085.

1c.18 National Guideline Clearinghouse or other URL: http://pediatrics.aappublications.org/content/pediatrics/129/4/e1085.full.pdf

1c.19 Grading of Strength of Guideline Recommendation. Has the recommendation been graded? Yes

1c.20 If guideline recommendation graded, identify the entity that graded the evidence including balance of representation and any disclosures regarding bias: Polin et al. grade all of the recommendations related to the strategies listed above and are too numerous to list here.

1c.21 System Used for Grading the Strength of Guideline Recommendation: CDC/Healthcare Infection Control Practice Advisory Committee System

1c.22 If other, identify and describe the grading scale with definitions:

1c.23 Grade Assigned to the Recommendation: .

1c.24 Rationale for Using this Guideline Over Others: Specificity to the NICU

Based on the NQF descriptions for rating the evidence, what was the developer's assessment of the quantity, quality, and consistency of the body of evidence?
1c.25 Quantity: High **1c.26** Quality: High**1c.27** Consistency: High

1. Evidence, Performance Gap, Priority – Importance to Measure and Report

Extent to which the specific measure focus is evidence-based, important to making significant gains in healthcare quality, and improving health outcomes for a specific high-priority (high-impact) aspect of healthcare where there is variation in or overall less-than-optimal performance. ***Measures must be judged to meet all subcriteria to pass this criterion and be evaluated against the remaining criteria.***

1a. Evidence to Support the Measure Focus – See attached Evidence Submission Form

Infection_evidence_Updated-635929648367835082.doc

1b. Performance Gap

Demonstration of quality problems and opportunity for improvement, i.e., data demonstrating:

- considerable variation, or overall less-than-optimal performance, in the quality of care across providers; and/or
- disparities in care across population groups.

1b.1. **Briefly explain the rationale for this measure** *(e.g., the benefits or improvements in quality envisioned by use of this measure)*
A bundle of improvement practices has been shown to dramatically reduce the frequency of hospital acquired infections for very low birth weight infants.

1b.2. **Provide performance scores on the measure as specified (current and over time) at the specified level of analysis.** *(This is required for endorsement maintenance. Include mean, std dev, min, max, interquartile range, scores by decile. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included). This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*
In 2014, 932 hospitals in the Vermont Oxford Network enrolled 56,099 very low birth weight infants of which 11.6% had a hospital acquired bacterial infection with an interquartile range among hospitals of 3.3% to 14.4%.

Year	N Hospitals		Mean	Minimum		Maximum		Q1	Q2	Q3
2006	672	0.192	0.000	1.000	0.086	0.170	0.264			
2007	724	0.18	0.000	1.000	0.086	0.169	0.249			
2008	799	0.172	0.000	1.000	0.079	0.153	0.231			
2009	839	0.157	0.000	1.000	0.077	0.136	0.213			
2010	789	0.139	0.000	1.000	0.063	0.115	0.191			
2011	793	0.132	0.000	1.000	0.061	0.111	0.179			
2012	910	0.12	0.000	1.000	0.048	0.102	0.166			
2013	921	0.112	0.000	0.677	0.042	0.095	0.156			
2014	938	0.108	0.000	0.813	0.038	0.089	0.148			

1b.3. **If no or limited performance data on the measure as specified is reported in 1b2, then provide a summary of data from the literature that indicates opportunity for improvement or overall less than optimal performance on the specific focus of measurement.**
Vermont Oxford Network VLBW Database Summary for 2014. Vermont Oxford Network. Burlington, VT. 2015.

1b.4. **Provide disparities data from the measure as specified (current and over time) by population group, e.g., by race/ethnicity, gender, age, insurance status, socioeconomic status, and/or disability.** *(This is required for endorsement maintenance. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities include.) This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*
Rates of hospital acquired bacterial infection varied by race/ethnicity of the mother, ranging from 11.4% for infants with black mothers to 8.9% for infants with Asian mothers.

Year	White non-Hispanic		Black non-Hispanic	Hispanic Asian
2006	0.195	0.234	0.215	0.176
2007	0.186	0.225	0.208	0.158
2008	0.178	0.201	0.192	0.143
2009	0.164	0.184	0.175	0.138
2010	0.147	0.160	0.160	0.125
2011	0.124	0.145	0.146	0.124
2012	0.123	0.139	0.127	0.104

2013	0.118	0.122	0.126	0.093
2014	0.116	0.126	0.125	0.109

1b.5. **If no or limited data on disparities from the measure as specified is reported in 1b4, then provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement. Include citations.**
Vermont Oxford Network VLBW Database Summary for 2014. Vermont Oxford Network. Burlington, VT. 2015.

1c. High Priority (previously referred to as High Impact)

The measure addresses:

- a specific national health goal/priority identified by DHHS or the National Priorities Partnership convened by NQF; OR
- a demonstrated high-priority (high-impact) aspect of healthcare (e.g., affects large numbers of patients and/or has a substantial impact for a smaller population; leading cause of morbidity/mortality; high resource use (current and/or future); severity of illness; and severity of patient/societal consequences of poor quality).

1c.1. **Demonstrated high priority aspect of healthcare**
Affects large numbers, A leading cause of morbidity/mortality, Patient/societal consequences of poor quality, Severity of illness
1c.2. **If Other:**

1c.3. **Provide epidemiologic or resource use data that demonstrates the measure addresses a high priority aspect of healthcare. List citations in 1c.4.**
Infants admitted to neonatal intensive care units are at high risk of hospital acquired infections. Hospital acquired infection in this population is associated with increased mortality, morbidity, length of stay and cost.

1c.4. **Citations for data demonstrating high priority provided in 1a.3**
Committee on Fetus and Newborn, Committee on Infectious Diseases. Epidemiology and diagnosis of health care-associated infections in the NICU. Pediatrics. 2012; 129:e1104.

1c.5. **If a PRO-PM (e.g. HRQoL/functional status, symptom/burden, experience with care, health-related behaviors), provide evidence that the target population values the measured PRO and finds it meaningful.** *(Describe how and from whom their input was obtained.)*

2. Reliability and Validity—Scientific Acceptability of Measure Properties

Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. ***Measures must be judged to meet the subcriteria for both reliability and validity to pass this criterion and be evaluated against the remaining criteria.***

2a.1. Specifications The measure is well defined and precisely specified so it can be implemented consistently within and across organizations and allows for comparability. eMeasures should be specified in the Health Quality Measures Format (HQMF) and the Quality Data Model (QDM).

De.5. **Subject/Topic Area** *(check all the areas that apply):*
Perinatal and Reproductive Health

De.6. **Cross Cutting Areas** *(check all the areas that apply):*
Safety : Healthcare Associated Infections

S.1. **Measure-specific Web Page** *(Provide a URL link to a web page specific for this measure that contains current detailed specifications including code lists, risk model details, and supplemental materials. Do not enter a URL linking to a home page or to general information.)*
https://public.vtoxford.org//wp-content/uploads/2014/03/NQF_Measure_0304.pdf

S.2a. **If this is an eMeasure**, HQMF specifications must be attached. Attach the zipped output from the eMeasure authoring tool (MAT) - if the MAT was not used, contact staff. (Use the specification fields in this online form for the plain-language description of the specifications)
This is not an eMeasure **Attachment:**

S.2b. **Data Dictionary, Code Table, or Value Sets** *(and risk model codes and coefficients when applicable) must be attached. (Excel or csv file in the suggested format preferred - if not, contact staff)*
Attachment **Attachment:** 0304_ICD_Code_Tables.xlsx

S.3. **For endorsement maintenance**, please briefly describe any changes to the measure specifications since last endorsement date and explain the reasons.
No changes

S.4. **Numerator Statement** *(Brief, narrative description of the measure focus or what is being measured about the target population, i.e., cases from the target population with the target process, condition, event, or outcome)*
IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
Eligible infants with one or more of the following criteria:
Criterion 1:
Bacterial Pathogen. A bacterial pathogen is recovered from a blood and/or cerebral spinal fluid culture obtained after Day 3 of life.
OR
Criterion 2:
Coagulase Negative Staphylococcus. The infant has all 3 of the following:
1. Coagulase negative staphylococcus is recovered from a blood culture obtained from either a central line, or peripheral blood sample and/or is recovered from cerebrospinal fluid obtained by lumbar puncture, ventricular tap or ventricular drain.
2. One or more signs of generalized infection (such as apnea, temperature instability, feeding intolerance, worsening respiratory distress or hemodynamic instability).
3. Teatment with 5 or more days of intravenous antibiotics after the above cultures were obtained. If the infant died, was discharged, or transferred prior to the completion of 5 days of intravenous antibiotics, this condition would still be met if the intention were to treat for 5 or more days.

S.5. **Time Period for Data** *(What is the time period in which data will be aggregated for the measure, e.g., 12 mo, 3 years, look back*

to August for flu vaccination? Note if there are different time periods for the numerator and denominator.)
Twelve months (Jan 1 - Dec 31)

S.6. **Numerator Details** *(All information required to identify and calculate the cases from the target population with the target process, condition, event, or outcome such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
IF an OUTCOME MEASURE, describe how the observed outcome is identified/counted. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
Infants whose birth weight is between 401 and 1500 grams or whose gestational age is between 22 weeks 0 days and 29 weeks 6 days are included if they have coagulase negative staphylococcus or one of the bacterial pathogens listed below after day 3 of life, provided they meet one of the following criteria:
1. They are born at the reporting hospital.
OR
2. They are admitted to any location in the reporting hospital within 28 days of birth, without first having gone home.
Bacterial Pathogens List:
1. Achromobacter species [including Achromobacter xylosoxidans (also known as Alcaligenes xylosoxidans) and others]
2. Acinetobacter species
3. Aeromonas species
4. Alcaligenes species [Alcaligenes xylosoxidans and others]
5. Bacteroides species
6. Burkholderia species [Burkholderia capecia and others]
7. Campylobacter species [Campylobacter fetus, C. jejuni and others]
8. Chryseobacterium species
9. Citrobacter species [Citrobacter diversus, C. freundii, C. koseri and others]
10. Clostridium species
11. Enterobacter species [Enterobacter aerogenes, E. cloacae, and others]
12. Enterococcus species [Enterococcus faecalis (also known as Streptococcus faecalis), E.faecium, and other Enterococcus species]
13. Escherichia coli
14. Flavobacterium species
15. Haemophilus species [Haemophilus influenzae and others]
16. Klebsiella species [Klebsiella oxytoca, K. pneumoniae and others]
17. Listeria monocytogenes
18. Moraxella species [Moraxella catarrhalis (also known as Branhamella catarrhalis) and others]
19. Neisseria species [Neisseria meningitidis, N. gonorrhoeae and others]
20. Pasteurella species
21. Prevotella species
22. Proteus species [Proteus mirabilis, P. vulgaris and others]
23. Providencia species [Providencia rettgeri, and others]
24. Pseudomonas species [Pseudomonas aeruginosa and others]
25. Ralstonia species
26. Salmonella species
27. Serratia species [Serratia liquefaciens, S. marcescens and others]
28. Staphylococcus coagulase positive [aureus]
29. Stenotrophomonas maltophilia
30. Streptococcus species [including Streptococcus Group A, Streptococcus Group B, Streptococcus Group D, Streptococcus pneumoniae, Strep milleri and others]

S.7. **Denominator Statement** *(Brief, narrative description of the target population being measured)*
Eligible infants who are in the reporting hospital after day 3 of life.

S.8. **Target Population Category** *(Check all the populations for which the measure is specified and tested if any):*
Children's Health, Populations at Risk : Dual eligible beneficiaries

S.9. **Denominator Details** *(All information required to identify and calculate the target population/denominator such as definitions, specific data collection items/responses , code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
Infants whose birth weights are between 401 and 1500 grams or whose gestational ages are between 22 weeks 0 days and 29 weeks 6 days are included if they are in the reporting hospital after day 3 of life, provided they meet one of the following criteria:
1. They are born at the reporting hospital.
OR
2. They are admitted to any location in the reporting hospital within 28 days of birth, without first having gone home.

S.10. **Denominator Exclusions** *(Brief narrative description of exclusions from the target population)*
Infants who do not meet eligibility criteria for birth weight, gestational age or hospital admission, or if the infant is discharged home, is transferred or dies prior to day 3 of life.

S.11. **Denominator Exclusion Details** *(All information required to identify and calculate exclusions from the denominator such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
1. Any infant who meets neither of the following conditions is excluded:
   - Birth weight between 401 and 1500 grams
   - Gestational age between 22 and 29 weeks.
2. Outborn infants who are admitted to the reporting hospital more than 28 days after birth are excluded.
3. Outborn infants who have been home prior to admission to the reporting hospital are excluded.
4. Infants discharged home on or before day 3 of life are excluded.
5. Infants who die on or before day 3 of life are excluded.
6. Infants who transfer to another hospital on or before day 3 of life and who are not readmitted to the reporting hospital.
7. Infants who transfer more than once prior to day 3 of life.

S.12. **Stratification Details/Variables** *(All information required to stratify the measure results including the stratification variables, definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format with at S.2b)*
N/A

S.13. **Risk Adjustment Type** (Select type. Provide specifications for risk stratification in S.12 and for statistical model in S.14-15)
Statistical risk model
If other:

S.14. **Identify the statistical risk model method and variables** *(Name the statistical method - e.g., logistic regression and list all the risk factor variables. Note - risk model development and testing should be addressed with measure testing under Scientific Acceptability)*
Logistic regression with shrinkage estimate - see S. 15a

S.15. **Detailed risk model specifications** *(must be in attached data dictionary/code list Excel or csv file. Also indicate if available at measure-specific URL identified in S.1.)*
Note: Risk model details (including coefficients, equations, codes with descriptors, definitions), should be provided on a separate worksheet in the suggested format in the Excel or csv file with data dictionary/code lists at S.2b.
Provided in response box S.15a

S.15a. **Detailed risk model specifications** (*if not provided in excel or csv file at S.2b*)
Logistic regression is used to model the infection measure on the following covariates: gestational age and its squared term; small for gestational age (Yes/No); multiple gestation (Yes/No); APGAR score at 1 minute (0-10); infant sex (Female, Male); vaginal delivery (Yes/No); major congenital malformation (Yes/No); and birth location (Inborn, Outborn). An estimate of the systematic variation associated with the hospital standardized morbidity ratios (SMRs) shrinks center SMR values and their confidence limits based on the number of infants reported (see, e.g., Martuzzi M and Hills M. Estimating the degree of heterogeneity between event rates using likelihood. American Journal of Epidemiology 1995; 141(4): 369-374; Simpson J et al. Analysing differences in clinical outcomes

between hospitals. Quality and Safety in Health Care 2003; (12): 257-262).

S.16. **Type of score:**
Other
If other: Standardized morbidity ratio and observed minus expected values with confidence bounds

S.17. **Interpretation of Score** *(Classifies interpretation of score according to whether better quality is associated with a higher score, a lower score, a score falling within a defined interval, or a passing score)*
Better quality = Lower score

S.18. **Calculation Algorithm/Measure Logic** *(Describe the calculation of the measure score as an ordered sequence of steps including identifying the target population; exclusions; cases meeting the target process, condition, event, or outcome; aggregating data; risk adjustment; etc.)*
1.Determine the number of infants for a reporting period who meet the population criteria described above. This number is termed N.

2.Using the definitions in the Network Manual of Operations, determine the number of infants who had nosocomial bacterial infection after day 3 of life and prior to discharge home for each of the N infants. This is the number of eligible infants who were diagnosed as having either coagulase negative staphylococcus and/or a late bacterial pathogen after day 3 of life. The number identified as having nosocomial bacterial infection is termed the "observed number with infection" or O for short.

3.For each of the N infants, calculate the expected value of infection by multiplying the coefficient times its covariate value for each covariate (coefficients provided on request). The covariates include:
Gestational Age in completed weeks (GA)
GA squared
Small for Gestational Age (data table provided on request)
Major birth defect (0=No, 1=Yes)
APGAR score at 1 minute (0 to 10)
Birth location (0=Inborn, 1=Outborn)
Multiple gestation (0=No, 1=Yes)
Infant gender (0=Female, 1=Male)
Mode of delivery (0=C-Section, 1=Vaginal)

4. Add the expected values for each of the N infants to calculate the number of expected cases of nosocomial bacterial infection. This number is termed the "expected number with infection" or E for short.

5. Calculate the standardized morbidity ratio (SMRshrnk) for nosocomial bacterial infection using the values for O and E and applying the estimate for systematic variation (v2), determined from Vermont Oxford Network analyses (provided on request).
SMRshrnk = (O + v2) / (E + v2)
with standard error SESMRshrnk=sqrt(1/(E+(1/v2)));
6. Calculate the shrunken, adjusted nosocomial bacterial infection rate
(Rateshrnk) and its 95% confidence interval.
Rateshrnk = (SMRshrnk x E) / N
with standard error (SERateshrnk) equal to SESMRshrnk x E) / N .
and 95% confidence interval for Rateshrnk equal to
Rateshrnk ± 1.96 × SERateshrnk.

7. Calculate the number of observed minus expected cases of nosocomial
bacterial infection, adjusting for case mix and systematic variation
(O–Eshrnk), and calculate the 95% control limits for O–Eshrnk.
O–Eshrnk = E / SMRshrnk
with 95% control limits equal to O–Eshrnk ± 1.96 × SESMRshrnk x E.

S.19. **Calculation Algorithm/Measure Logic Diagram URL or Attachment** *(You also may provide a diagram of the Calculation Algorithm/Measure Logic described above at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
URL

S.20. **Sampling** *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample*

size.)
IF a PRO-PM, identify whether (and how) proxy responses are allowed.
Data for all eligible infants born during the reporting period are collected.

S.21. **Survey/Patient-reported data** *(If measure is based on a survey, provide instructions for conducting the survey and guidance on minimum response rate.)*
IF a PRO-PM, specify calculation of response rates to be reported with performance measure results.

S.22. **Missing data** (specify how missing data are handled, e.g., imputation, delete case.)
Required for Composites and PRO-PMs.
Cases missing outcome data are deleted.

S.23. **Data Source** *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*
If other, please describe in S.24.
Electronic Clinical Data : Registry

S.24. **Data Source or Collection Instrument** *(Identify the specific data source/data collection instrument e.g. name of database, clinical registry, collection instrument, etc.)*
IF a PRO-PM, identify the specific PROM(s); and standard methods, modes, and languages of administration.
Vermont Oxford Network Database

S.25. **Data Source or Collection Instrument** *(available at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
No data collection instrument provided

S.26. **Level of Analysis** *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*
Facility

S.27. **Care Setting** *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*
Hospital/Acute Care Facility
If other:

S.28. **COMPOSITE Performance Measure** - Additional Specifications *(Use this section as needed for aggregation and weighting rules, or calculation of individual performance measures if not individually endorsed.)*

NATIONAL QUALITY FORUM—Measure Testing (subcriteria 2a2, 2b2-2b7)

Measure Number (*if previously endorsed*)**:** 0304

Measure Title: Late sepsis or meningitis in Very Low Birth Weight (VLBW) neonates (risk-adjusted)

Date of Submission: 2/15/2016

Type of Measure:

☐ Composite – ***STOP – use composite testing form***	☒ Outcome (*including PRO-PM*)
☐ Cost/resource	☐ Process
☐ Efficiency	☐ Structure

Instructions

- Measures must be tested for all the data sources and levels of analyses that are specified. ***If there is more than one set of data specifications or more than one level of analysis, contact NQF staff*** about how to

present all the testing information in one form.

- **For <u>all</u> measures, sections 1, 2a2, 2b2, 2b3, and 2b5 must be completed.**
- **For <u>outcome and resource use</u> measures**, section **2b4** also must be completed.
- If specified for **<u>multiple data sources/sets of specificaitons</u>** (e.g., claims and EHRs), section **2b6** also must be completed.
- Respond to <u>all</u> questions as instructed with answers immediately following the question. All information on testing to demonstrate meeting the subcriteria for reliability (2a2) and validity (2b2-2b6) must be in this form. An appendix for *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 20 pages (*incuding questions/instructions;* minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.
- For information on the most updated guidance on how to address sociodemographic variables and testing in this form refer to the release notes for version 6.6 of the Measure Testing Attachment.

<u>Note</u>: The information provided in this form is intended to aid the Steering Committee and other stakeholders in understanding to what degree the testing results for this measure meet NQF's evaluation criteria for testing.

2a2. Reliability testing [10] demonstrates the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise. For **PRO-PMs and composite performance measures**, reliability should be demonstrated for the computed performance score.

2b2. Validity testing [11] demonstrates that the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality. For **PRO-PMs and composite performance measures**, validity should be demonstrated for the computed performance score.

2b3. Exclusions are supported by the clinical evidence; otherwise, they are supported by evidence of sufficient frequency of occurrence so that results are distorted without the exclusion; [12]
AND
If patient preference (e.g., informed decisionmaking) is a basis for exclusion, there must be evidence that the exclusion impacts performance on the measure; in such cases, the measure must be specified so that the information about patient preference and the effect on the measure is transparent (e.g., numerator category computed separately, denominator exclusion category computed separately). [13]

2b4. For outcome measures and other measures when indicated (e.g., resource use):

- **an evidence-based risk-adjustment strategy** (e.g., risk models, risk stratification) is specified; is based on patient factors (including clinical and sociodemographic factors) that influence the measured outcome and are present at start of care; [14,15] and has demonstrated adequate discrimination and calibration

OR

- rationale/data support no risk adjustment/ stratification.

2b5. Data analysis of computed measure scores demonstrates that methods for scoring and analysis of the specified measure allow for **identification of statistically significant and practically/clinically meaningful** [16] **differences in performance**;
OR
there is evidence of overall less-than-optimal performance.

2b6. If multiple data sources/methods are specified, there is demonstration they produce comparable results.

2b7. For **eMeasures, composites, and PRO-PMs** (or other measures susceptible to missing data), analyses identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not

biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias.

Notes

10. Reliability testing applies to both the data elements and computed measure score. Examples of reliability testing for data elements include, but are not limited to: inter-rater/abstractor or intra-rater/abstractor studies; internal consistency for multi-item scales; test-retest for survey items. Reliability testing of the measure score addresses precision of measurement (e.g., signal-to-noise).

11. Validity testing applies to both the data elements and computed measure score. Validity testing of data elements typically analyzes agreement with another authoritative source of the same information. Examples of validity testing of the measure score include, but are not limited to: testing hypotheses that the measures scores indicate quality of care, e.g., measure scores are different for groups known to have differences in quality assessed by another valid quality measure or method; correlation of measure scores with another valid indicator of quality for the specific topic; or relationship to conceptually related measures (e.g., scores on process measures to scores on outcome measures). Face validity of the measure score as a quality indicator may be adequate if accomplished through a systematic and transparent process, by identified experts, and explicitly addresses whether performance scores resulting from the measure as specified can be used to distinguish good from poor quality.

12. Examples of evidence that an exclusion distorts measure results include, but are not limited to: frequency of occurrence, variability of exclusions across providers, and sensitivity analyses with and without the exclusion.

13. Patient preference is not a clinical exception to eligibility and can be influenced by provider interventions.

14. Risk factors that influence outcomes should not be specified as exclusions

15. With large enough sample sizes, small differences that are statistically significant may or may not be practically or clinically meaningful. The substantive question may be, for example, whether a statistically significant difference of one percentage point in the percentage of patients who received smoking cessation counseling (e.g., 74 percent v. 75 percent) is clinically meaningful; or whether a statistically significant difference of $25 in cost for an episode of care (e.g., $5,000 v. $5,025) is practically meaningful. Measures with overall less-than-optimal performance may not demonstrate much variability across providers.

1. DATA/SAMPLE USED FOR ALL TESTING OF THIS MEASURE

Often the same data are used for all aspects of measure testing. In an effort to eliminate duplication, the first five questions apply to all measure testing. If there are differences by aspect of testing,(e.g., reliability vs. validity) be sure to indicate the specific differences in question 1.7.

1.1. What type of data was used for testing? (*Check all the sources of data identified in the measure specifications and data used for testing the measure. Testing must be provided for all the sources of data specified and intended for measure implementation.* ***If different data sources are used for the numerator and denominator, indicate N [numerator] or D [denominator] after the checkbox.***)

Measure Specified to Use Data From: (*must be consistent with data sources entered in S.23*)	**Measure Tested with Data From:**
☐ abstracted from paper record	☐ abstracted from paper record
☐ administrative claims	☐ administrative claims
☒ clinical database/registry	☒ clinical database/registry
☐ abstracted from electronic health record	☐ abstracted from electronic health record
☐ eMeasure (HQMF) implemented in EHRs	☐ eMeasure (HQMF) implemented in EHRs
☐ other: Click here to describe	☐ other: Click here to describe

1.2. If an existing dataset was used, identify the specific dataset (*the dataset used for testing must be consistent with the measure specifications for target population and healthcare entities being measured; e.g., Medicare Part A claims, Medicaid claims, other commercial insurance, nursing home MDS, home health OASIS, clinical registry*). Vermont Oxford Network Database

1.3. What are the dates of the data used in testing? 2006 to 2014

1.4. What levels of analysis were tested? (*testing must be provided for all the levels specified and intended for measure implementation, e.g., individual clinician, hospital, health plan*)

Measure Specified to Measure Performance of: (*must be consistent with levels entered in item S.26*)	**Measure Tested at Level of:**
☐ individual clinician	☐ individual clinician
☐ group/practice	☐ group/practice
☒ hospital/facility/agency	☒ hospital/facility/agency
☐ health plan	☐ health plan
☐ other:	☐ other:

1.5. How many and which measured entities were included in the testing and analysis (by level of analysis and data source)? (*identify the number and descriptive characteristics of measured entities included in the analysis (e.g., size, location, type); if a sample was used, describe how entities were selected for inclusion in the sample*) 1,044 hospitals contributed at least one year of data.

1.6. How many and which patients were included in the testing and analysis (by level of analysis and data source)? (*identify the number and descriptive characteristics of patients included in the analysis (e.g., age, sex, race, diagnosis); if a sample was used, describe how patients were selected for inclusion in the sample*) Over the nine year period, 474,131 infants who were 501 to 1500 grams at birth and survived to three days of life were registered with Vermont Oxford Network.

1.7. If there are differences in the data or sample used for different aspects of testing (e.g., reliability, validity, exclusions, risk adjustment), identify how the data or sample are different for each aspect of testing reported below. None

1.8 What were the patient-level sociodemographic (SDS) variables that were available and analyzed in the data or sample used? For example, patient-reported data (e.g., income, education, language), proxy variables when SDS data are not collected from each patient (e.g. census tract), or patient community characteristics (e.g. percent vacant housing, crime rate). Vermont Oxford Network does not collect SDS data.

2a2. RELIABILITY TESTING

***Note**: If accuracy/correctness (validity) of data elements was empirically tested, separate reliability testing of data elements is not required – in 2a2.1 check critical data elements; in 2a2.2 enter "see section 2b2 for validity testing of data elements"; and skip 2a2.3 and 2a2.4.*

2a2.1. What level of reliability testing was conducted? (*may be one or both levels*)

☐ **Critical data elements used in the measure** (*e.g., inter-abstractor reliability; data element reliability must address ALL critical data elements*)

☒ **Performance measure score** (e.g., *signal-to-noise analysis*)

2a2.2. For each level checked above, describe the method of reliability testing and what it tests (*describe the steps—do not just name a method; what type of error does it test; what statistical analysis was used*) Logistic regression models are tested for performance using the area under the receiver operating characteristic curve and monitored over time.

We did a split-half analysis to measure the internal consistency of the measure. We selected hospitals with at least 10 infants, ran 100 samples of patients within each hospital randomly split to two groups, computed hospital rates, and calculated the correlations.

2a2.3. For each level of testing checked above, what were the statistical results from reliability testing? (e.*g., percent agreement and kappa for the critical data elements; distribution of reliability statistics from a signal-to-noise analysis*)
Correlation coefficients are:

2006	0.62
2007	0.62
2008	0.61
2009	0.61
2010	0.58
2011	0.56
2012	0.56
2013	0.58
2014	0.63

2a2.4 What is your interpretation of the results in terms of demonstrating reliability? (i.*e., what do the results mean and what are the norms for the test conducted?*)
The correlation coefficients were lower than we expected. It suggests that the definition may not be applied in the same manner across all infants at all hospitals. We did notice that the coefficients increased as the number of infants at the hospitals increased.

2b2. VALIDITY TESTING

2b2.1. What level of validity testing was conducted? (*may be one or both levels*)

☐ **Critical data elements** (*data element validity must address ALL critical data elements*)

☐ **Performance measure score**

- ☐ **Empirical validity testing**
- ☒ **Systematic assessment of face validity of performance measure score as an indicator** of quality or resource use

(*i.e., is an accurate reflection of performance on quality or resource use and can distinguish good from poor performance*)

2b2.2. For each level of testing checked above, describe the method of validity testing and what it tests (*describe the steps—do not just name a method; what was tested, e.g., accuracy of data elements compared to authoritative source, relationship to another measure as expected; what statistical analysis was used)* The measure, including the list of bacterial pathogens, was developed by board certified neonatologists and reviewed by clinical experts in neonatal infection. The measure is reviewed annually by the Vermont Oxford Network Database Advisory Committee, consisting of national and international experts in the neonatal community. The bacterial pathogens list was last revised in 2015. Comprehensive business rules test reach record for consistency, completeness and accuracy. Submitted records with errors must be corrected before data are finalized and reports of the measure are provided to hospitals.

2b2.3. What were the statistical results from validity testing? (*e.g., correlation; t-test*) No testing done

2b2.4. What is your interpretation of the results in terms of demonstrating validity? (i.*e., what do the results mean and what are the norms for the test conducted?*) N/A

_____

2b3. EXCLUSIONS ANALYSIS

NA ☐ no exclusions — *skip to section 2b4*

2b3.1. Describe the method of testing exclusions and what it tests (*describe the steps—do not just name a method; what was tested, e.g., whether exclusions affect overall performance scores; what statistical analysis was used*) The population measured includes premature infants with birth weights between 501 and 1500 grams. The occurrence of infection is monitored after day 3 of life while the infant is hospitalized in the reporting hospital. For infants who transfer to another hospital, monitoring continues when the infant is readmitted to the reporting hospital. Infants who are discharged home are no longer monitored. These rules for tracking infants provide a reasonably homogenous population base for performance inferences and quality improvement decisions.

Business rules require that infants be hospitalized more than three days or the measure is not applicable. Other exclusions are also enforced by business rules that assure database integrity.

2b3.2. What were the statistical results from testing exclusions? (*include overall number and percentage of individuals excluded, frequency distribution of exclusions across measured entities, and impact on performance measure scores*)

Year	Number of infants	Number of infants excluded	% excluded
2006	41657	1921	4.6
2007	49866	2196	4.4
2008	52799	2151	4.1
2009	54042	2275	4.2
2010	53500	2052	3.8
2011	55170	2180	4.0
2012	55638	2110	3.8
2013	56168	2152	3.8
2014	56730	2186	3.9

2b3.3. What is your interpretation of the results in terms of demonstrating that exclusions are needed to prevent unfair distortion of performance results? (*i.e., the value outweighs the burden of increased data collection and analysis. Note: **If patient preference is an exclusion**, the measure must be specified so that the effect on the performance score is*

transparent, e.g., scores with and without exclusion) "Exclusion" was defined as having length of stay <3 days. Given that the definition of the measure is onset of infection >3 days, such exclusions are necessary.

2b4. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES
If not an intermediate or health outcome, or PRO-PM, or resource use measure, skip to section 2b5.

2b4.1. What method of controlling for differences in case mix is used?
☐ **No risk adjustment or stratification**
☒ **Statistical risk model with** 8 **risk factors**
☐ **Stratification by** Click here to enter number of categories **risk categories**
☐ **Other,** Click here to enter description

2b4.2. If an outcome or resource use measure is not risk adjusted or stratified, provide rationale and analyses to demonstrate that controlling for differences in patient characteristics (case mix) is not needed to achieve fair comparisons across measured entities.

2b4.3. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors (clinical factors or sociodemographic factors) used in the statistical risk model or for stratification by risk (*e.g., potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of p<0.10; correlation of x or higher; patient factors should be present at the start of care*)

The number of hospitals and number of patients vary by birth year based on the number of hospitals participating in the registry. Each reporting hospital submits data for all eligible infants as described in the Specifications section of this submission form. The number of hospitals for the period 2006-2010 is tabulated in item 2b2.1. above. The number of infection cases for this period is tabulated in item 2b3.3. above (Number of Infants minus Number Excluded).

The logistic regression model includes the risk factors listed below.
- Gestational age in completed weeks and its quadratic term.
- Small for gestational age (SGA, Yes or No), defined as being in the 10th percentile or less for birth weight, given the infant's gestational age, the maternal race, the infant's gender and whether the infant was a singleton or multiple gestation. The United States Natality Datasets are used for calculating the 10th percentile values.
- Major birth defect (Yes or No).
- Multiple gestation (Yes or No).
- APGAR score at 1 minute (0 to 10).
- Infant gender (Male or Female).
- Maternal race (Hispanic, White, Asian or Other - Black is the reference category).
- Vaginal delivery (Yes or No).
- Birth location (Inborn or Outborn).

When one or more predictor variables is missing for infants with a known outcome measure, an imputation procedure is used based on Network or center specific rates for the missing values.
The adjusted rates are "shrunken" to remove random variation in signals of performance using an empirical Bayesian method. For an example of this method, see Martuzzi M and Hills M, Estimating the degree of heterogeneity between event rates using likelihood, Am J of Epi, 141, 4, 369-374 (1995) and Simpson J et al, Analysing differences in clinical outcomes between hospitals, Qual Saf
Health Care, 12, 257-262 (2003).

The model contains characteristics of the infants at birth to adjust for differences in case mix upon admission to the neonatal ICU. The model specifically does not include process of care measures since those can differ from hospital to hospital.

2b4.4a. What were the statistical results of the analyses used to select risk factors?

For infants born in 2010, the coefficients with standard errors and chi-square values are listed below. These values are consistents with values obtained for infants born in previous years.

Parameter	DF	Estimate	Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-7.9290	1.7004	21.7427	<.0001
GAWeeks	1	0.8195	0.1248	43.1333	<.0001
GASQ	1	-0.0208	0.00228	83.4953	<.0001
Male	1	0.1015	0.0309	10.8079	0.0010
MultipleBirth	1	0.0164	0.0357	0.2094	0.6472
Vaginal	1	-0.0140	0.0337	0.1718	0.6785
BirthDefect	1	0.5357	0.0898	35.5700	<.0001
SmallForGA	1	0.5490	0.0514	114.2164	<.0001
AP1	1	-0.0208	0.00700	8.7861	0.0030
HispRace	1	0.0677	0.0495	1.8692	0.1716
WhiteRace	1	-0.0667	0.0378	3.1091	0.0779
AsianRace	1	-0.1357	0.0845	2.5765	0.1085
OthRace	1	0.1283	0.0979	1.7177	0.1900
Outborn	1	0.1838	0.0400	21.1029	<.0001

Characteristics were selected *a priori* without statistical differentiation. We do review the coefficients for the models by year to check for year-to-year stability.

2b4.4b. Describe the analyses and interpretation resulting in the decision to select SDS factors (e.g. prevalence of the factor across measured entities, empirical association with the outcome, contribution of unique variation in the outcome, assessment of between-unit effects and within-unit effects) Vermont Oxford Network does not collect SDS measures.

2b4.5. Describe the method of testing/analysis used to develop and validate the adequacy of the statistical model or stratification approach (*describe the steps—do not just name a method; what statistical analysis was used*) Hosmer-Lemeshow goodness of fit statistics

Provide the statistical results from testing the approach to controlling for differences in patient characteristics (case mix) below.

If stratified, skip to 2b4.9

2b4.6. Statistical Risk Model Discrimination Statistics (*e.g., c-statistic, R-squared*)**:**

2b4.7. Statistical Risk Model Calibration Statistics (*e.g., Hosmer-Lemeshow statistic*):

Birth Year	Fit Chi-Square
2006	26.0
2007	41.9
2008	18.0
2009	37.8
2010	30.6

2011 35.4
2012 14.8
2013 17.9
2014 13.5

2b4.8. Statistical Risk Model Calibration – Risk decile plots or calibration curves:

2b4.9. Results of Risk Stratification Analysis:

2b4.10. What is your interpretation of the results in terms of demonstrating adequacy of controlling for differences in patient characteristics (case mix)? (i.*e., what do the results mean and what are the norms for the test conducted*) The model used to include race and ethnicity as a proxy for SDS. We removed these variables in 2012. The model fit improved since that time.

2b4.11. Optional Additional Testing for Risk Adjustment (*not required, but would provide additional support of adequacy of risk model, e.g., testing of risk model in another data set; sensitivity analysis for missing data; other methods that were assessed*)

2b5. IDENTIFICATION OF STATISTICALLY SIGNIFICANT & MEANINGFUL DIFFERENCES IN PERFORMANCE

2b5.1. Describe the method for determining if statistically significant and clinically/practically meaningful differences in performance measure scores among the measured entities can be identified (*describe the steps—do not just name a method; what statistical analysis was used? Do not just repeat the information provided related to performance gap in 1b)*

If the 95% upper bound for the standardized morbidity ratio (SMR - observed value / expected value) is less than one, the hospital performance is classified as "better than expected"; if the 95% lower bound for the SMR is greater than one, the hospital performance is classified as "worse than expected"; if the 95% lower and upper bounds for the SMR inlcudes one, performance is classified as "as expected." An estimate of the systematic variation associated with the hospital standardized morbidity ratios (SMRs) and the observed minus expected measure is used to shrink center SMR values and their confidence limits based on the number of infants reported. The shrinkage method is the gamma-Poisson approach to filtering random variation associated with infection as a risk adjusted indicator of performance.

2b5.2. What were the statistical results from testing the ability to identify statistically significant and/or clinically/practically meaningful differences in performance measure scores across measured entities? (e.g., *number and percentage of entities with scores that were statistically significantly different from mean or some benchmark, different from expected; how was meaningful difference defined*) Of the 932 eligible hospitals in 2014, 770 performed as expected (i.e., the confidence interval contained 1), 92 performed worse than expected (i.e., the lower bound was greater than 1), and 70 performed better than expected (i.e., the upper bound was less than 1).

2b5.3. What is your interpretation of the results in terms of demonstrating the ability to identify statistically significant and/or clinically/practically meaningful differences in performance across measured entities? (i.*e., what do the results mean in terms of statistical and meaningful differences?*) Given the shrinkage factor, we expect a number of hospitals to perform as expected with outliers at the tails. Hospitals can use the SMRs or observed minus expected to understand whether bacterial infection is of concern for their hospitals after adjustment. Hospitals receive annual guidance on how to interpret the measures.

2b6. COMPARABILITY OF PERFORMANCE SCORES WHEN MORE THAN ONE SET OF SPECIFICATIONS

If only one set of specifications, this section can be skipped.

Note: *This item is directed to measures that are risk-adjusted (with or without SDS factors)* ***OR*** *to measures with more than one set of specifications/instructions (e.g., one set of specifications for how to identify and compute the measure*

from medical record abstraction and a different set of specifications for claims or eMeasures). It does not apply to measures that use more than one source of data in one set of specifications/instructions (e.g., claims data to identify the denominator and medical record abstraction for the numerator). ***Comparability is not required when comparing performance scores with and without SDS factors in the risk adjustment model. However, if comparability is not demonstrated for measures with more than one set of specifications/instructions, the different specifications (e.g., for medical records vs. claims) should be submitted as separate measures.***

2b6.1. Describe the method of testing conducted to compare performance scores for the same entities across the different data sources/specifications (*describe the steps—do not just name a method; what statistical analysis was used*)

2b6.2. What were the statistical results from testing comparability of performance scores for the same entities when using different data sources/specifications? (*e.g., correlation, rank order*)

2b6.3. What is your interpretation of the results in terms of the differences in performance measure scores for the same entities across the different data sources/specifications? (i.*e., what do the results mean and what are the norms for the test conducted*)

2b7. MISSING DATA ANALYSIS AND MINIMIZING BIAS

2b7.1. Describe the method of testing conducted to identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias (*describe the steps—do not just name a method; what statistical analysis was used*) Not done

2b7.2. What is the overall frequency of missing data, the distribution of missing data across providers, and the results from testing related to missing data? (*e.g., results of sensitivity analysis of the effect of various rules for missing data/nonresponse; if no empirical sensitivity analysis, identify the approaches for handling missing data that were considered and pros and cons of each*) Over the nine years, 378 infants (0.08%) were missing data. Missing data does not appear to be an issue.

2b7.3. What is your interpretation of the results in terms of demonstrating that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias**?** (i.*e., what do the results mean in terms of supporting the selected approach for missing data and what are the norms for the test conducted; if no empirical analysis, provide rationale for the selected approach for missing data*)

3. Feasibility

Extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

3a. Byproduct of Care Processes

For clinical measures, the required data elements are routinely generated and used during care delivery (e.g., blood pressure, lab test, diagnosis, medication order).

3a.1. **Data Elements Generated as Byproduct of Care Processes.**

generated by and used by healthcare personnel during the provision of care, e.g., blood pressure, lab value, medical condition, Coded by someone other than person obtaining original information (e.g., DRG, ICD-9 codes on claims), Abstracted from a record by someone other than person obtaining original information (e.g., chart abstraction for quality measure or registry)

If other:

3b. Electronic Sources

The required data elements are available in electronic health records or other electronic sources. If the required data are not in electronic health records or existing electronic sources, a credible, near-term path to electronic collection is specified.

3b.1. **To what extent are the specified data elements available electronically in defined fields?** *(i.e., data elements that are needed to compute the performance measure score are in defined, computer-readable fields)*

ALL data elements are in defined fields in a combination of electronic sources

3b.2. **If ALL the data elements needed to compute the performance measure score are not from electronic sources, specify a credible, near-term path to electronic capture, OR provide a rationale for using other than electronic sources.**

3b.3. **If this is an eMeasure, provide a summary of the feasibility assessment in an attached file or make available at a measure-specific URL.**

No feasibility assessment **Attachment:**

3c. Data Collection Strategy

Demonstration that the data collection strategy (e.g., source, timing, frequency, sampling, patient confidentiality, costs associated with fees/licensing of proprietary measures) can be implemented (e.g., already in operational use, or testing demonstrates that it is ready to put into operational use). For eMeasures, a feasibility assessment addresses the data elements and measure logic and demonstrates the eMeasure can be implemented or feasibility concerns can be adequately addressed.

3c.1. **Describe what you have learned/modified as a result of testing and/or operational use of the measure regarding data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, time and cost of data collection, other feasibility/implementation issues.**

IF a PRO-PM, consider implications for both individuals providing PROM data (patients, service recipients, respondents) and those whose performance is being measured.

Patient identifiers are not collected in the registry. Confidentiality for each hospital member is strictly maintained. Procedures in place assure reasonable confidence that data are complete and accurate. There are no specific fees for this measure, although members of the Vermont Oxford Network pay an annual membership fee.

3c.2. **Describe any fees, licensing, or other requirements to use any aspect of the measure as specified** *(e.g., value/code set, risk model, programming code, algorithm)*.

4. Usability and Use

Extent to which potential audiences (e.g., consumers, purchasers, providers, policy makers) are using or could use performance results for both accountability and performance improvement to achieve the goal of high-quality, efficient healthcare for individuals or populations.

4a. Accountability and Transparency

Performance results are used in at least one accountability application within three years after initial endorsement and are publicly reported within six years after initial endorsement (or the data on performance results are available). If not in use at the time of initial endorsement, then a credible plan for implementation within the specified timeframes is provided.

4.1. **Current and Planned Use**

NQF-endorsed measures are expected to be used in at least one accountability application within 3 years and publicly reported within 6 years of initial endorsement in addition to performance improvement.

Planned	Current Use (for current use provide URL)
Public Reporting	Quality Improvement with Benchmarking (external benchmarking to multiple organizations) Vermont Oxford Network Password protected website Quality Improvement (Internal to the specific organization) N/A N/A

4a.1. **For each CURRENT use, checked above, provide:**

- Name of program and sponsor
- Purpose
- Geographic area and number and percentage of accountable entities and patients included

Vermont Oxford Network has nearly 1000 members. We do not know what proportion are using this measure for quality improvement.

4a.2. **If not currently publicly reported OR used in at least one other accountability application (e.g., payment program, certification, licensing) what are the reasons?** (*e.g., Do policies or actions of the developer/steward or accountable entities restrict access to performance results or impede implementation?*)

There are no publicly-reported quality measures for the neonatal population. Vermont Oxford Network members use reports provided by the Network to do internal quality improvement and benchmarking. Additionally, Vermont Oxford Network has managed several quality improvement collaboratives for its members on reducing infection.

4a.3. **If not currently publicly reported OR used in at least one other accountability application, provide a credible plan for implementation within the expected timeframes -- any accountability application within 3 years and publicly reported within 6 years of initial endorsement.** (*Credible plan includes the specific program, purpose, intended audience, and timeline for implementing the measure within the specified timeframes. A plan for accountability applications addresses mechanisms for data aggregation and reporting.*)

Vermont Oxford Network is committed to working with accrediting bodies that are developing publicly-reported quality measures for the neonatal population.

4b. Improvement

Progress toward achieving the goal of high-quality, efficient healthcare for individuals or populations is demonstrated. If not in use for performance improvement at the time of initial endorsement, then a credible rationale describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.

4b.1. **Progress on Improvement. (Not required for initial endorsement unless available.)**

Performance results on this measure (current and over time) should be provided in 1b.2 and 1b.4. Discuss:

- **Progress (trends in performance results, number and percentage of people receiving high-quality healthcare)**
- **Geographic area and number and percentage of accountable entities and patients included**

From 2007 to 2013, rates of any late infection at Vermont Oxford Network member hospitals in the United States continued to

decrease for all infants. Overall, 78% of hospitals reported decreases during this time period. (Horbar JD, Edwards EM, Soll RF, Edwards WH, Buus-Frank M. Beyond CLABSI: late onset infection rates for infants with birth weights of 401 to 1500 grams from 2007 to 2013. Pediatric Academic Society 2015 Annual Meeting.)

4b.2. **If no improvement was demonstrated, what are the reasons? If not in use for performance improvement at the time of initial endorsement, provide a credible rationale that describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.**

4c. Unintended Consequences

The benefits of the performance measure in facilitating progress toward achieving high-quality, efficient healthcare for individuals or populations outweigh evidence of unintended negative consequences to individuals or populations (if such evidence exists).

4c.1. **Were any unintended negative consequences to individuals or populations identified during testing; OR has evidence of unintended negative consequences to individuals or populations been reported since implementation? If so, identify the negative unintended consequences and describe how benefits outweigh them or actions taken to mitigate them.**

To mitigate unintended consequences associated with Vermont Oxford Network member hospitals reporting nosocomial bacterial infection, members receive a manual of operations annually that contains definitions and clearly operationalized criteria for the measure. Comprehensive business rules verify records for consistency, completeness and accuracy. Centers employ a definitive process to assure that the measure is not reported until data are complete and correct. Hospital contacts must verify that data for all eligible infants are submitted prior to finalization.

5. Comparison to Related or Competing Measures

If a measure meets the above criteria and there are endorsed or new related measures (either the same measure focus or the same target population) or competing measures (both the same measure focus and the same target population), the measures are compared to address harmonization and/or selection of the best measure.

5. Relation to Other NQF-endorsed Measures
Are there related measures (conceptually, either same measure focus or target population) or competing measures (conceptually both the same measure focus and same target population)? If yes, list the NQF # and title of all related and/or competing measures.
Yes

5.1a. **List of related or competing measures (selected from NQF-endorsed measures)**
0478 : Neonatal Blood Stream Infection Rate (NQI 03)
1731 : PC-04 Health Care-Associated Bloodstream Infections in Newborns

5.1b. **If related or competing measures are not NQF endorsed please indicate measure title and steward.**

5a. Harmonization

The measure specifications are harmonized with related measures;
OR
The differences in specifications are justified

5a.1. **If this measure conceptually addresses EITHER the same measure focus OR the same target population as NQF-endorsed measure(s):**
Are the measure specifications completely harmonized?
No

5a.2. **If the measure specifications are not completely harmonized, identify the differences, rationale, and impact on interpretability and data collection burden.**
The target populations are different, as are the item definitions and risk adjustment methodology.

5b. Competing Measures

The measure is superior to competing measures (e.g., is a more valid or efficient way to measure);
OR
Multiple measures are justified.

5b.1. **If this measure conceptually addresses both the same measure focus and the same target population as NQF-endorsed measure(s):**
Describe why this measure is superior to competing measures (e.g., a more valid or efficient way to measure quality); OR provide a rationale for the additive value of endorsing an additional measure. (Provide analyses when possible.)
N/A

Appendix

A.1 Supplemental materials may be provided in an appendix. All supplemental materials (such as data collection instrument or methodology reports) should be organized in one file with a table of contents or bookmarks. If material pertains to a specific submission form number, that should be indicated. Requested information should be provided in the submission form and required attachments. There is no guarantee that supplemental materials will be reviewed.

Attachment:

Contact Information

Co.1 **Measure Steward (Intellectual Property Owner):** Vermont Oxford Network
Co.2 **Point of Contact:** Erika, Edwards, eedwards@vtoxford.org, 802-865-4814-246
Co.3 **Measure Developer if different from Measure Steward:** Vermont Oxford Network

Co.4 **Point of Contact:** Erika, Edwards, eedwards@vtoxford.org, 802-865-4814-246

Additional Information

Ad.1 **Workgroup/Expert Panel involved in measure development**
Provide a list of sponsoring organizations and workgroup/panel members' names and organizations. Describe the members' role in measure development.
N/A

Measure Developer/Steward Updates and Ongoing Maintenance
Ad.2 **Year the measure was first released:** 2008
Ad.3 **Month and Year of most recent revision:** 02, 2016
Ad.4 **What is your frequency for review/update of this measure?** Annual
Ad.5 **When is the next scheduled review/update for this measure?** 02, 2017

Ad.6 **Copyright statement:** Copyright © 2016 Vermont Oxford Network, Inc.
Ad.7 **Disclaimers:**

Ad.8 **Additional Information/Comments:**

MEASURE WORKSHEET

This document summarizes the evaluation of the measure as it progresses through NQF's Consensus Development Process (CDP). The information submitted by measure developers/stewards is included after the Brief Measure Information, Preliminary Analysis, and Pre-meeting Public and Member Comments sections.

To navigate the links in the worksheet: Ctrl + click link to go to the link; ALT + LEFT ARROW to return

Brief Measure Information

NQF #: 0469
Measure Title: PC-01 Elective Delivery
Measure Steward: The Joint Commission
Brief Description of Measure: This measure assesses patients with elective vaginal deliveries or elective cesarean births at >= 37 and < 39 weeks of gestation completed. This measure is a part of a set of five nationally implemented measures that address perinatal care (PC-02: Cesarean Birth, PC-03: Antenatal Steroids, PC-04: Health Care-Associated Bloodstream Infections in Newborns, PC-05: Exclusive Breast Milk Feeding)
Developer Rationale: A reduction in the number of non-medically indicated elective deliveries at >=37 to <39 weeks gestation will result in a substantial decrease in neonatal morbidity and mortality, as well as a significant savings in health care costs. In addition, the rate of cesarean sections should decrease with fewer elective inductions resulting in decreased length of stay and health care costs.

The measure will assist health care organizations (HCOs) to track non-medically indicated early term elective deliveries and reduce the occurrence.

Numerator Statement: Patients with elective deliveries with ICD-10-PCS Principal Procedure Code or ICD-10-PCS Other Procedure Codes for one or more of the following:
• Medical induction of labor as defined in Appendix A, Table 11.05 available at: http://manual.jointcommission.org/releases/TJC2015B2/ while not in Labor prior to the procedure
• Cesarean birth as defined in Appendix A, Table 11.06 and all of the following:
not in Labor
no history of a Prior Uterine Surgery available at: http://manual.jointcommission.org/releases/TJC2015B2/
Denominator Statement: Patients delivering newborns with >= 37 and < 39 weeks of gestation completed with ICD-10-PCS Principal or Other Procedure Codes for delivery as defined in Appendix A, Table 11.01.1 available at: http://manual.jointcommission.org/releases/TJC2015B2/ and with ICD-10-CM Principal Diagnosis Code or ICD-10-CM Other Diagnosis Codes for planned cesarean birth in labor as defined in Appendix A, Table 11.06.1 available at: http://manual.jointcommission.org/releases/TJC2015B2/
Denominator Exclusions: • ICD-10-CM Principal Diagnosis Code or ICD-10-CM Other Diagnosis Codes for conditions possibly justifying elective delivery prior to 39 weeks gestation as defined in Appendix A, Table 11.07
• Less than 8 years of age
• Greater than or equal to 65 years of age
• Length of Stay >120 days
• Enrolled in clinical trials
• Gestational Age < 37 or >= 39 weeks or UTD

Measure Type: Process
Data Source: Electronic Clinical Data, Paper Medical Records
Level of Analysis: Facility, Population : National

IF Endorsement Maintenance – Original Endorsement Date: Oct 24, 2008 **Most Recent Endorsement Date:** Apr 02, 2012

Maintenance of Endorsement -- Preliminary Analysis

To maintain NQF endorsement endorsed measures are evaluated periodically to ensure that the measures still meets the NQF endorsement criteria ("maintenance"). The emphasis for maintaining endorsement is focused on how effective the measure is for promoting improvements in quality. Endorsed measures should have some experience from the field to inform the evaluation. The emphasis for maintaining endorsement is noted for each criterion.

Criteria 1: Importance to Measure and Report

1a. Evidence
Maintenance measures – less emphasis on evidence unless there is new information or change in evidence since the prior evaluation.

1a. Evidence. The evidence requirements for a *process or intermediate outcome* measure is that it is based on a systematic review (SR) and grading of the body of empirical evidence where the specific focus of the evidence matches what is being measured.

The developer provides the following evidence for this measure:

- **Systematic Review of the evidence specific to this measure?** ☒ **Yes** ☐ **No**
- **Quality, Quantity and Consistency of evidence provided?** ☒ **Yes** ☐ **No**
- **Evidence graded?** ☒ **Yes** ☐ **No**

Summary of prior review in 2012

- The developer reported that "both ACOG and AAP have had guidelines in place for a number of years which do not support non-medically indicated elective deliveries at< 39 weeks gestation. Several studies consistently document increased morbidity associated with elective delivery before 39 weeks. The studies note that elective deliveries performed at < 39 weeks carry significant risk for the newborn (odds ratios 2.0-3-0 compared to newborns born between 39-41 weeks). " The previous Perinatal Committee agreed that the "evidence is strong that elective delivery prior to 39 weeks impacts newborn adversely."

Changes to evidence from last review

☒ **The developer attests that there have been no changes in the evidence since the measure was last evaluated.**
☐ **The developer provided updated evidence for this measure:**

Updates:

- In 2013 ACOG reaffirmed the Practice Bulletin for Induction of Labor (August 2009).
- An ACOG Committee Opinion from April 2013 (reaffirmed in 2015) reviewed the evidence and provided guidance for "Nonmedically Indicated Early-Term Deliveries" concluding that "Although there are specific indications for delivery before 39 weeks of gestation, a nonmedically indicated early-term delivery is not appropriate."

Exception to evidence NA

Questions for the Committee:

- *The developer attests the underlying evidence for the measure has not changed since the last NQF endorsement review. Does the Committee agree the evidence basis for the measure has not changed and there is no need for repeat discussion and voting on Evidence?*

Preliminary rating for evidence: ☒ **Pass** ☐ **No Pass**

1b. Gap in Care/Opportunity for Improvement and 1b. Disparities
Maintenance measures – increased emphasis on gap and variation

1b. Performance Gap. The performance gap requirements include demonstrating quality problems and opportunity for improvement.

The developers provide data for the past several years:
Baseline 2Q 2010: 164 hospitals, 11,843 cases; national aggregate rate = 18.8%; 50th %tile = 15.5%
Beginning 2014, all hospitals with >110 births/year were required to report.

	2011	2012	2013	2014
# hospitals	166	170	200	1388
# patients	13,907	13,404	14,880	130,882
National aggregate rate	13.6%	8.0%	4.4%	3.3%
10th-25th -90th %tile	1.5 -5 -9.8– 31.5%	0– 2.6-4.9-21.2%	0- 0- 2.6-13.9%	0-0 - 2.1 -8.7 %
Mean hospital rate (SD)	0.13998 (0.13183)	0.08296 (0.09555)	0.05737 (0.10193)	0.82(0.31)

Beginning in January 2016 hospitals with >300 births/year are required to report – an additional 821 (approximately 80% of all birthing hospitals).

Disparities
Data from use of this measure to understand any disparities is not provided. The developers refer to the literature that has reported increased rates of elective inductions for all races though higher rates are seen in non-Hispanic whites. Bailey, et al. (2014) recently reviewed data for 638 rural women, recruited prenatally from three counties in rural southern Appalachia, who delivered electively at = 37 weeks. Those delivered electively at early term were 7.7 times more likely to be low birth weight, 4.4 times more likely to have a neonatal intensive care unit admission, and 2.5 times more likely to develop jaundice. Patients living furthest from the hospital were most likely to deliver electively at <39 weeks.

Questions for the Committee:
- *Since this a "lower equals better" measure, is 0% realistic? Is there room for a decrease beyond 3.3% nationally?*
- *The measure results decreased in 2014 even though 1188 more hospitals reported on the measure. Do you expect the additional 821 hospitals reporting in 2016 to have poorer performance than those currently reporting?*
- *Are there specific populations for which the performance might be less good?*
- *Are there potential patient safety concerns addressed by this measure regardless of the high performance rates?*
- *Is this measure useful in understanding disparities in provision of healthcare to pregnant women?*
- *Is there a gap in care that warrants a national performance measure?*

Preliminary rating for opportunity for improvement: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments
Criteria 1: Importance to Measure and Report (including 1a, 1b, 1c)

1a. Evidence to Support Measure Focus
Comments:
Evidence is directly related to this process measure and can lead to adverse events.
If anything the evidence has improved to say that stillbirths did not go up due to the "39 week" rule.

1b. Performance Gap
Comments:
Performance data was provided. Still have less than optimal performance nationally, although much improvement has been seen in the last several years. Data related to disparities related to subgroups is not provided.

0% is not realistic but there is still variation that can be brought down. I would suspect that the hospitals not reporting do not perform well on this measure. No reason to think that this measure performs differently on different groups Safety concerns are a potential increase in stillbirths, see comment above. Ironically, "poor quality" care for this measure typically effects high socioeconomic women. ie a reverse disparity.

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability

2a1. Reliability Specifications
Maintenance measures – no change in emphasis – specifications should be evaluated the same as with new measures

2a1. Specifications requires the measure, as specified, to produce consistent (reliable) and credible (valid) results about the quality of care when implemented.

Data source(s): Electronic Clinical Data, Paper Medical Records; Vital Records reports, delivery logs and clinical information systems were added as acceptable data sources to help hospitals identify all cases with 37 and 38 weeks gestation

Specifications:

- The numerator captures all medical inductions of labor (not in labor) and Cesarean sections (not in labor) and without prior uterine surgery (four data elements).
- The denominator captures all deliveries between 37 0/7 and 38 6/7 weeks (seven data elements.)
- Exclusions include conditions for possibly justifying elective delivery.

Changes from the previous specifications include:

- All ICD-9-CM diagnosis codes and ICD-9-CM procedure codes were converted to ICD-10-CM diagnosis and ICD-10-PCS procedure codes throughout the measure specifications,
- The initial patient population is now identified with ICD-10-PCS-Principal or Other Procedure Codes for delivery instead of diagnosis codes for pregnancy, since the ICD-10-CM Principal or Other Diagnosis Codes do not indicate whether the delivery took place during the hospitalization.
- The numerator included population for medical induction of labor now requires a check for the presence of labor prior to the procedure, since the ICD-10-PCS-Principal and Other Procedure Codes assigned for labor induction do not distinguish between induction and augmentation of labor.
- A new numerator data element Prior Uterine Surgery was added for cases with cesarean births in order to pass cases with acceptable prior uterine surgeries as noted below under S.6 numerator details.
- New ICD-10-CM Diagnosis Codes were added to Table 11.07: Conditions Possibly Justifying Elective Delivery Prior to 39 Weeks Gestation in Appendix A.
- A calculation algorithm is described. Sampling is allowed.

Questions for the Committee:

- *Are all the data elements clearly defined? Are all appropriate codes included?*
- *Are the changes to the conditions possibly justifying elective delivery appropriate?*
- *Is the logic or calculation algorithm clear?*
- *Is it likely this measure can be consistently implemented?*

2a2. Reliability Testing Testing attachment
Maintenance measures – less emphasis if no new testing data provided

2a2. Reliability testing demonstrates if the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise enough to distinguish differences in performance across providers.

For maintenance measures, summarize the reliability testing from the prior review:

Inter-rater reliability (IRR) was performed by ORXY vendor reabstaraction for 108 hospitals comprising 13,279 records. IRR

is an appropriate method of assessing data element reliability for chart abstraction. The agreement rate for the data elements "Active labor" was 94.29% and "Gestational age" was 89.75%.

Describe any updates to testing none

SUMMARY OF TESTING

Reliability testing level ☐ **Measure score** ☒ **Data element** ☐ **Both**

Reliability testing performed with the data source and level of analysis indicated for this measure ☒ **Yes** ☐ **No**

Method(s) of reliability testing see above

Results of reliability testing see above Only %agreement -- no statistical test such as Kappa or ICC was provided.

Guidance from the Reliability Algorithm

Precise specifications (Box 1) → Empirical reliability testing (Box 2) → Data element testing (Box 8) → appropriate method- IRR (Box 9) → high or moderate confidence of reliability of numerator data element → moderate (highest rating possible)

Questions for the Committee:

- *Were all critical data elements tested?*
- *Is the test sample adequate to generalize for widespread implementation?*
- *Do the results demonstrate sufficient reliability so that differences in performance can be identified?*

Preliminary rating for reliability: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

2b. Validity
Maintenance measures – less emphasis if no new testing data provided

2b1. Validity: Specifications

2b1. Validity Specifications. This section should determine if the measure specifications are consistent with the evidence.

Specifications consistent with evidence in 1a. ☒ **Yes** ☐ **Somewhat** ☐ **No**
Specification not completely consistent with evidence

Question for the Committee:

- *Are the specifications consistent with the evidence?*
- *Are the exclusions evidence-based?*

2b2. Validity testing

2b2. Validity Testing should demonstrate the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality.

For maintenance measures, summarize the validity testing from the prior review: face validity only

Describe any updates to validity testing: new empirical validity testing of the measure score is provided

SUMMARY OF TESTING

Validity testing level ☒ **Measure score** ☐ **Data element testing against a gold standard** ☐ **Both**

Method of validity testing of the measure score:

- ☐ **Face validity only**
- ☒ **Empirical validity testing of the measure score**

Validity testing method:

Using one year of data from 2014-2015 for 1345 hospitals and 2,695,467 patient records, the Spearman rank-order correlation (a nonparametric measure of association based on the ranks of the data values by measure PC-01 and hospitals) was used to correlate the results from this measure with other measures in the Joint Commission's perinatal set. The developer hypothesizes that hospitals that perform well on this measure will perform will on all measures in the set.

Validity testing results: The developer provides scatter plots and a correlation table and summarized the findings: "The correlation of PC-01 with the other PC measures in the PC measure set indicates that the correlations (with the exception of PC-05), although in the expected direction and statistically significant, are relatively weak. Although 90% of the hospital measure rates fall between 0 and 6.7%, there are still a number of hospitals with measure rates significantly greater than 6.7%, indicating that the performance of hospitals on this measure are not uniformly acceptable."

Questions for the Committee:

- *The correlations among the five measures are relatively weak. DO you expect stronger correlations of performance?*
- *Is the test sample adequate to generalize for widespread implementation?*
- *Do the results demonstrate sufficient validity so that conclusions about quality can be made?*
- *Do you agree that the score from this measure as specified is an indicator of quality?*

2b3-2b7. Threats to Validity

2b3. Exclusions:

Frequency of exclusions were presented:

Data element	Overall %	Median%	Min - Max
Clinical Trials	0.07	0	0 – 6.59%
Gestational age <37 and >39 weeks	84.9%	75.2%	0.29 – 28%

Questions for the Committee:

- *Are the exclusions consistent with the evidence?*
- *Are any patients or patient groups inappropriately excluded from the measure?*
- *Are the exclusions/exceptions of sufficient frequency and variation across providers to be needed (and outweigh the data collection burden)?*

2b4. Risk adjustment: **Risk-adjustment method** ☒ **None** ☐ **Statistical model** ☐ **Stratification**

2b5. Meaningful difference *(can statistically significant and clinically/practically meaningful differences in performance measure scores can be identified)*:

The developer provides the following descriptive statistics for the measure results:

N = 1,237 hospitals n = 1,130,083
Mean: 2.74%
Min = 0%
Percentile 10%: 0%
Percentile 25%: 0%
Median: 1.7%
Percentile 75%: 3.7%
Percentile 90%: 6.7%
Max = 51.2%

The Joint Commission's Target Analysis uses two methods: Given the national average for a performance measure, a target range is constructed. Using generalized linear mixed models methodology (also known as hierarchical models), a predicted estimate of an HCO's performance, with a corresponding 95% confidence interval, is generated. This confidence interval is compared to the target range, to determine the HCOs' rating. The estimate of the organization's

true performance is based on both the data from that organization and on data from the entire set of reporting organizations. The object of target analysis is to compare a health care organizations (HCO) data against a comparative norm for the purpose of evaluating performance improvement opportunities. When an organization's performance level is statistically significantly different from a comparative norm, it is considered a statistical deviation. A statistical deviation may be desirable or undesirable depending on the "direction of improvement" of the measure.

PC-01 Distribution of Outliers 2011 1st Quarter Data:
Scores on this measure: N=160, Mean 13.6%, SD 0.1594
10th Percentile= 0%
25th Percentile= 0%
50th Percentile= 9%
75th Percentile= 19%
90th Percentile= 34%

156 (97.5%) Neutral – results not significantly different from target range
4 (2.5%) Unfavorable - results statistically significantly lower than the national rate

Question for the Committee:

- *Does this measure identify meaningful differences about quality?*

2b6. Comparability of data sources/methods: NA

2b7. Missing Data
The developer reports that "Any file with missing data will result in a measure category assignment of X and rejection of the file from the warehouse, unless the data are not used to process the measure. If the data are used to process the measure and have been reported by the abstractor as No/UTD, the case will result in a measure category assignment of D (i.e., failed measure, not rejected)."

Guidance from algorithm: Specifications consistent with evidence (Box 1) → Assessed all potential threats to validity (Box 2) **Meaningful differences may be an issue here** → empirical testing (Box 3) →testing of measure score (Box 6) →appropriate testing method (Box 7) → testing results moderate (or possible low) → moderate

Preliminary rating for validity: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 2: Scientific Acceptability of Measure Properties (including all 2a, 2b, and 2d)

2a1. & 2b1. Specifications

Comments:

Clearly defined elements. Codes are provided, however there may be additional codes that could be included but it is near impossible to include them all.
Elements are clearly defines. Changes are appropriate with ICD10. Logic is clear and measure can be consistently implemented. Critical elements were tested.
Specifications are consistent with the evidence.
Exclusions are appropriate.

Algorithm is clear.
No concerns regarding consistent implementation.

2a2. Reliability Testing

Comments:

Adequate scope. Agreement rate was reported but not statistical test.
*I do not expect correlations between measures. Test sample is adequate. THis is an appropriate measure of quality.**

2b2. Validity Testing
Comments:
Adequate scope using Spearman correlation. High rates are linked to poor outcomes in this population.
Yes.

2b3. Exclusions Analysis
2b4. Risk Adjustment/Stratification for Outcome or Resource Use Measures
2b5. Identification of Statistically Significant & Meaningful Differences In Performance
2b6. Comparability of Performance Scores When More Than One Set of Specifications
2b7. Missing Data Analysis and Minimizing Bias
Comments:
No, these are not a large percentage of the population.
Missing data is not a threat to validity. Sampling can be done if hospital infrastructure does not allow this to be done from ICD10 codes.

Criterion 3. Feasibility
Maintenance measures – no change in emphasis – implementation issues may be more prominent

3. Feasibility is the extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

- According to the developer "hospitals using this performance measure generally collect measure data via manual review of the paper medical record, the EMR or a combination of both." and
- "not all hospitals currently have the capacity to abstract the electronic version of this measure, so continues to offer this chart abstracted version which allows for data capture from unstructured data fields. All data elements needed to compute the PC-01 performance measure score have been retooled for capture from electronic sources. "

Questions for the Committee:
- *What is the burden for hospitals in collecting data for this measure?*
- *Are the required data elements routinely generated and used during care delivery?*
- *What is the burden of data collection for this measure?*
- *Are the required data elements available in electronic form, e.g., EHR or other electronic sources?*

Preliminary rating for feasibility: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments
Criteria 3: Feasibility

3a. Byproduct of Care Processes
3b. Electronic Sources
3c. Data Collection Strategy
Comments:
Routinely generated. Available electronically. Concerns are regarding the process of QA for coding that needs to occur with accurate data for this measure.
Burden to collect is not huge, hospitals are already doing it so to change would be a greater burden. Data elements are basic and routinely used.

Criterion 4: Usability and Use
Maintenance measures – increased emphasis – much greater focus on measure use and usefulness, including both

impact /improvement and unintended consequences

4. Usability and Use evaluate the extent to which audiences (e.g., consumers, purchasers, providers, policymakers) use or could use performance results for both accountability and performance improvement activities.

Current uses of the measure
Publicly reported? ☒ **Yes** ☐ **No**

Current use in an accountability program? ☒ **Yes** ☐ **No**
OR
Planned use in an accountability program? ☐ **Yes** ☐ **No**

Accountability program details

- Public Reporting --Quality Check® http://www.qualitycheck.org/consumer/searchQCR.aspx
- Regulatory and Accreditation Programs Hospital Accreditation Program http://jointcommission.org
- Hospital Compare (CMS) https://www.medicare.gov/hospitalcompare/search.html
- Hospital Inpatient Quality Reporting Program https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/HospitalQualityInits/HospitalRHQDAPU.html
- Hospital Value Based Purchasing Program https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/hospital-value-based-purchasing/index.html?redirect=/hospital-value-based-purchasing/

Improvement results
The developer reports: "The rate of early term elective deliveries has decreased from 18.8% in 2010 with 164 hospitals reporting to 3.3% in 2014 with 1388 hospitals reporting based on Joint Commission ORYX performance measurement data. Beginning with January 1, 2016 discharges, an additional 821 accredited hospitals will begin reporting the data. The new reporting requirement will capture approximately 80% of the accredited hospitals with maternity services in the US. Additionally, the adoption of PC-01 into the CMS Hospital Inpatient Quality Reporting (HIQR) Program has significantly increased the number of hospitals reporting their early term elective delivery data with 3500 hospitals participating in the reporting program."

Unexpected findings (positive or negative) during implementation
The developers report on three issues that were addressed since onset of implementation:

- Cases with prior uterine surgery were inappropriately failing the measure – ACTION: The measure rate calculation algorithm was revised to include a check prior to a cesarean birth via a new data element (Prior Uterine Surgery) created to enable cases with prior uterine surgery to remain in the denominator population and pass the measure.
- Patients who did not receive prenatal care were inappropriately included in the measure denominator, as the gestational age data element was abstracted as unable to be determined (UTD) – ACTION: In order to avoid penalizing hospitals, cases with UTD were removed from the measure population.
- Some hospitals have reported higher rates due to small denominator populations as a result of sampling – ACTION: Vital Records reports, delivery logs and clinical information systems were added as acceptable data sources to help hospitals identify all cases with 37 and 38 weeks gestation, so that 100% of these cases could be reviewed to increase the denominator population size.

Potential harms see above

Feedback: MAP supported this measures for Hospital Inpatient Quality Reporting (2012) and Hospital Value Based Purchasing (2013)

***Questions for the Committee**:*

- *Committee members who are using /have used this measure should share their experiences. Are you aware of any*

unintended consequences of this measure? Are you seeing lower C-section rates and/or fewer NICU admissions?

- *Are the revisions to the measure resulting from the mitigating actions appropriate?*
- *How can the performance results be used to further the goal of high-quality, efficient healthcare?*
- *Do the benefits of the measure outweigh any potential unintended consequences?*

Preliminary rating for usability and use: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 4: Usability and Use

4a. Accountability and Transparency

4b. Improvement

4c. Unintended Consequences

Comments:

TJC Perinatal Core Measure, Quality Check, CMS, Value Based Purchasing. Need to set a lower goal for the nation. 5% is very doable and we can do better. No unintended consequences in my opinion except for an impact on patient satisfaction if proper scripting from the providers regarding health and safety does not occur.

**Revisions are appropriate. major concern is for increased stillbirth and or people using measure to justify inappropriately keeping high risk pregnancies continuing. Stillbirth issue has new evidence suggesting no/not a large effect on stillbirth. Benefit most likely outweighs harms. **

Criterion 5: Related and Competing Measures

Related or competing measures

Measure 2829 is a new eMeasure version of this measure.

Harmonization

[Harmonization]

-

Pre-meeting public and member comments

-

1. IMPACT, OPPORTUITY, EVIDENCE - IMPORTANCE TO MEASURE AND REPORT

Importance to Measure and Report is a threshold criterion that must be met in order to recommend a measure for endorsement. All three subcriteria must be met to pass this criterion. See guidance on evidence.

Measures must be judged to be important to measure and report in order to be evaluated against the remaining criteria. (evaluation criteria)

1c.1 Structure-Process-Outcome Relationship *(Briefly state the measure focus, e.g., health outcome, intermediate clinical outcome, process, structure; then identify the appropriate links, e.g., structure-process-health outcome; process- health outcome; intermediate clinical outcome-health outcome)*:

The focus of the measure is to decrease the number of elective deliveries >> population determined >> population assessed >> patient delivers spontaneously or planned delivery greater or equal to 39 weeks gestation >> improved maternal and fetal outcomes >> decreased length of stay and fetal morbidity and mortality.

1c.2-3 Type of Evidence *(Check all that apply)*:

Clinical Practice Guideline

Systematic review of body of evidence (other than within guideline development)

1c.4 Directness of Evidence to the Specified Measure *(State the central topic, population, and outcomes addressed in the body of evidence and identify any differences from the measure focus and measure target population)*:

The central topic for the measure is the reduction of elective deliveries at >= 37 and < 39 weeks of gestation completed. The target population for the performance measure is consistent with the body of evidence supporting the reduction of elective deliveries.

1c.5 Quantity of Studies in the Body of Evidence *(Total number of studies, not articles)*: No randomized-control trials (RCTs) were identified for early-term elective deliveries. RCTs were only identified for post-term elective deliveries versus expectant management. Given the current amount of population data available on the harms of early term and late pre-term delivery, it would be unethical to conduct such a study. Several studies were identified which were retrospective cohort or prospective observational in design examining thousands of births and the potential for adverse outcomes for both mother and newborn. In addition, several recent studies were identified addressing quality improvement interventions that were successful in reducing non-medically indicated early term elective deliveries.

1c.6 Quality of Body of Evidence (*Summarize the certainty or confidence in the estimates of benefits and harms to patients across studies in the body of evidence resulting from study factors. Please address: a) study design/flaws; b) directness/indirectness of the evidence to this measure (e.g., interventions, comparisons, outcomes assessed, population included in the evidence); and c) imprecision/wide confidence intervals due to few patients or events):* The quality of evidence supporting the reduction in the number of non-medically indicated elective deliveries is moderate . It is noteworthy to examine the fact that randomized control trials cannot be conducted, as one cannot randomly select women to agree to an elective delivery at < 39 weeks gestation.

As previously noted, both ACOG and AAP have had guidelines in place for a number of years which do not support non-medically indicated elective deliveries at > 39 weeks gestation. Several studies consistently document increased morbidity associated with elective delivery before 39 weeks. The studies note that elective deliveries performed at < 39 weeks carry significant risk for the newborn (odds ratios 2.0-3-0 compared to newborns born between 39-41 weeks).

In spite of the fact that all studies reviewed were either retrospective or prospective cohort studies, no study design flaws were noted.

1c.7 Consistency of Results across Studies *(Summarize the consistency of the magnitude and direction of the effect):* The body of evidence consistently supports the benefit of reduction of non-medically indicated early term elective deliveries. All studies show an increase in the number of neonatal morbidities associated with early term deliveries, subsequent reduction of elective non-medically indicated deliveries reduces harm to the neonate. All studies demonstrated similar findings related to the direction of effect, though the magnitude varied from study to study, i.e., 8-17.8% increase in NICU admissions, rates of adverse respiratory outcomes, mechanical ventilation, newborn sepsis, hypoglycemia, admission to the NICU and hospitalization of 5 days or more increased by a factor of 1.8 to 4.2. and the incidence of transient tachypnea of the newborn, respiratory distress syndrome (RDS) and persistent pulmonary hypertension of the newborn were 3.1%, 0.25% and .17% respectively.

1c.8 Net Benefit *(Provide estimates of effect for benefit/outcome; identify harms addressed and estimates of effect; and net benefit -*

benefit over harms):
As described before, elective deliveries performed at =>39 weeks gestation results in improved maternal and neonatal outcomes and will result in a substantial decrease in cesarean sections and neonatal morbidity, as well as substantial savings in health care costs. A recent study showed that by waiting until 39 weeks gestation, the NICU admissions fell from 12.8% to 5.9%, RDS fell from 3.7% to 0.9%, newborn sepsis fell from 7.0% to 2.5% and hospitalization > 5 days fell from 9.1% to 3.6%. This same study estimated that one-half million newborn intensive care unit days could be avoided in the U.S. population were a national rate of 1.7% to be achieved, with cost savings approaching $1 billion annually.

1c.9 **Grading of Strength/Quality of the Body of Evidence**. Has the body of evidence been graded? No

1c.10 **If body of evidence graded, identify the entity that graded the evidence including balance of representation and any disclosures regarding bias:** Not Applicable

1c.11 **System Used for Grading the Body of Evidence:** Other

1c.12 **If other, identify and describe the grading scale with definitions:** Although grading of the evidence was not determined during our systematic review, it was determined that the guideline developers accounted for a balanced representation of information, looked beyond one specialty group or discipline, and provided information that was accessible and met the requirements set out in this measure maintenance form.

1c.13 **Grade Assigned to the Body of Evidence:** Not Applicable

1c.14 **Summary of Controversy/Contradictory Evidence:** There is no documented evidence regarding controversy related to the reduction of non-medically indicated early term elective deliveries. A review of recent studies also supports the use of quality improvement interventions to further reduce the number of such deliveries.

1c.15 **Citations for Evidence other than Guidelines*(Guidelines addressed below)*:**

- American Academy of Family Physicians. (2000). Tips from Other Journals: Elective induction doubles cesarean delivery rate, 61, 4.Retrieved September 16, 2011 at: http://www.aafp.org/afp/20000215/tips/39.html.
- American College of Obstetricians and Gynecologists. (November 1996). ACOG Educational Bulletin.
- American College of Obstetricians and Gynecologists (ACOG) Practice Bulletin No 107. (2009). Induction of labor. Obstetrics & Gynecology. 114(2). 386-97.
- Clark, S., Miller, D., Belfort, M., Dildy, G., Frye, D., & Meyers, J. (2009). Neonatal and maternal outcomes associated with elective delivery. [Electronic Version]. Am J Obstet Gynecol. 200:156.e1-156.e4.
- Clark, S., Frye, D., Meyers, J., Belfort, M., Dildy, G., Kofford, S et al. (2010). Reduction in elective delivery at <39 weeks of gestation: comparative effectiveness of 3 approaches to change and the impact on neonatal intensive care admission and stillbirth. Am J Obstet Gynecol. 203:449.e1-6.
- Davidoff, M., Dias, T., Damus, K., Russell, R., Bettegowda, V.R., Dolan, S., et al. (2006). Changes in the gestational age distributin among U.S. singleton births: impacts on rates of late preterm birth, 1992-2002. Semin Perinatol. Feb;30(1):8-15.
- Engle, W.A. & Kominiarek, M.A. (2008). Late preterm infants, early term infants, and timing of elective deliveries. Clin Perinatol. 35:325-41.
- Glantz, J. (Apr.2005). Elective induction vs. spontaneous labor associations and outcomes. [Electronic Version]. J Reprod Med. 50(4):235-40.
- Tita, A., Landon, M., Spong, C., Lai, Y., Leveno, K., Varner, M, et al. (2009). Timing of elective repeat cesarean delivery at term and neonatal outcomes. [Electronic Version]. NEJM. 360:2, 111-120.

1c.16 **Quote verbatim, <u>the specific guideline recommendation</u>** *(Including guideline # and/or page #)*:
Clinical Considerations and Recommendations Page 389

What are the indications and contraindications to induction of labor?

Indications for induction of labor are not absolute but should take into account maternal and fetal conditions, gestational age, cervical status, and other factors. Following are examples of maternal or fetal conditions that may be indications for induction of labor:
- Abruptio placentae
- Chorioamnionitis
- Fetal demise

• Gestational hypertension
• Preeclampsia, eclampsia
• Premature rupture of membranes
• Postterm pregnancy
• Maternal medical conditions (eg, diabetes mellitus, renal disease, chronic pulmonary disease, chronic hypertension, antiphospholipid syndrome)
• Fetal compromise (eg, severe fetal growth restriction, isoimmunization, oligohydramnios)

Labor also may be induced for logistic reasons, for example, risk of rapid labor, distance from hospital, or psychosocial indications. In such circumstances, at least one of the gestational age criteria in the box should be met, or fetal lung maturity should be established. A mature fetal lung test result before 39 weeks of gestation, in the absence of appropriate clinical circumstances, is not an indication for delivery. The individual patient and clinical situation should be considered in determining when induction of labor is contraindicated. Generally, the contraindications to labor induction are the same as those for spontaneous labor and vaginal delivery. They include, but are not limited to, the following situations:
• Vasa previa or complete placenta previa
• Transverse fetal lie
• Umbilical cord prolapse
• Previous classical cesarean delivery
• Active genital herpes infection
• Previous myomectomy entering the endometrial cavity

What criteria should be met before the cervix is ripened or labor is induced?

Assessment of gestational age and consideration of any potential risks to the mother or fetus are of paramount importance for appropriate evaluation and counseling before initiating cervical ripening or labor induction. The patient should be counseled regarding the indications for induction, the agents and methods of labor stimulation, and the possible need for repeat induction or cesarean delivery. Although prospective studies are limited in evaluating the benefits of elective induction of labor, nulliparous women undergoing induction of labor with unfavorable cervices should be counseled about a twofold increased risk of cesarean delivery (Level II-2). In addition, labor progression differs significantly for women with an elective induction of labor compared with women who have spontaneous onset of labor (Level II-2).
Allowing at least 12–18 hours of latent labor before diagnosing a failed induction may reduce the risk of cesarean delivery (Level II-2, 3). Additional requirements for cervical ripening and induction of labor include assessment of the cervix, pelvis, fetal size, and presentation. Monitoring FHR and uterine contractions is recommended as for any high-risk patient in active labor. Although trained nursing personnel can monitor labor induction, a physician capable of performing a cesarean delivery should be readily available.

1c.17 Clinical Practice Guideline Citation: • American College of Obstetricians and Gynecologists (ACOG) Practice Bulletin No 107. (2009). Induction of labor. Obstetrics & Gynecology. 114(2). 386-97.

1c.18 National Guideline Clearinghouse or other URL:

1c.19 Grading of Strength of Guideline Recommendation. Has the recommendation been graded? Yes

1c.20 If guideline recommendation graded, identify the entity that graded the evidence including balance of representation and any disclosures regarding bias: The American College of Obstericians and Gynecologists

1c.21 System Used for Grading the Strength of Guideline Recommendation: USPSTF

1c.22 If other, identify and describe the grading scale with definitions:

1c.23 Grade Assigned to the Recommendation: Level II

1c.24 Rationale for Using this Guideline Over Others: The American College of Obstetricians and Gynecologists the nation´s leading group of professionals providing health care for women. Practice Bulletins provide obstetricians and gynecologists with current information on established techniques and clinical management guidelines. The American College of Obstetricians and Gynecologists (the College) continuously surveys the field for advances to be incorporated in these series and monitors existing bulletins to ensure they are current. Individual bulletins are withdrawn from and added to the series on a continuing basis and reaffirmed periodically.

Based on the NQF descriptions for rating the evidence, what was the <u>developer's assessment</u> of the quantity, quality, and consistency of the body of evidence?

1c.25 Quantity: High **1c.26** Quality: Moderate**1c.27** Consistency**:** High

1. Evidence, Performance Gap, Priority – Importance to Measure and Report

Extent to which the specific measure focus is evidence-based, important to making significant gains in healthcare quality, and improving health outcomes for a specific high-priority (high-impact) aspect of healthcare where there is variation in or overall less-than-optimal performance. ***Measures must be judged to meet all subcriteria to pass this criterion and be evaluated against the remaining criteria.***

1a. Evidence to Support the Measure Focus – See attached Evidence Submission Form
0469_Evidence_MSF5.0_Data.doc

1b. Performance Gap

Demonstration of quality problems and opportunity for improvement, i.e., data demonstrating:

- considerable variation, or overall less-than-optimal performance, in the quality of care across providers; and/or
- disparities in care across population groups.

1b.1. **Briefly explain the rationale for this measure** *(e.g., the benefits or improvements in quality envisioned by use of this measure)*
A reduction in the number of non-medically indicated elective deliveries at >=37 to <39 weeks gestation will result in a substantial decrease in neonatal morbidity and mortality, as well as a significant savings in health care costs. In addition, the rate of cesarean sections should decrease with fewer elective inductions resulting in decreased length of stay and health care costs.

The measure will assist health care organizations (HCOs) to track non-medically indicated early term elective deliveries and reduce the occurrence.

1b.2. **Provide performance scores on the measure as specified (current and over time) at the specified level of analysis.** *(This is required for endorsement maintenance. Include mean, std dev, min, max, interquartile range, scores by decile. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included). This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*
Early term elective deliveries are still being performed; however, the performance gap has narrowed over time. A goal of 5% or less based on recommendations from the PC Technical Advisory Panel (TAP) should be achievable. The Perinatal Care (PC) core measures were added as a new core measure set in 2010 for hospitals to select in order to meet their ORYX performance measurement requirement for Joint Commission accreditation purposes. At that time, approximately 164 hospitals reported the data with an average measure rate of 18.8% (n=11,843 patients). In January 2014, The Joint Commission required mandatory reporting of the PC measure set for all accredited hospitals with 1100 births or more annually. 1388 hospitals reported the data with an average rate of 3.4% (n=130,882 patients). It is important to note that a performance gap of 3.7% exists for the 90th percentile of hospitals performing at 8.7% (if 5% is considered goal performance). The threshold for mandatory reporting was recently lowered to 300 births annually effective January 2016. The new reporting requirement will now capture approximately 80% of all accredited birthing hospitals. As a result, the rates may increase with the addition of approximately 821 more hospitals reporting data. Below is the specified level of analysis for PC-01 beginning with discharges April 1, 2010 through December 31, 2014.
2Q 2010: 11,843 denominator cases; 2,231 numerator cases; 164 hospitals; 18.8% national aggregate rate; 0.17827 mean of hospital rates; 0.12745 standard deviation; 33.3% 90th percentile rate; 23.7% 75th percentile rate/upper quartile; 15.5% 50th percentile rate/median rate; 9.0% 25th percentile rate/lower quartile; and 4.7% 10th percentile rate.
CY 2011: 1,3907 denominator cases; 1,892 numerator cases; 166 hospitals; 13.6% national aggregate rate; 0.13998 mean of hospital rates; 0.13183 standard deviation; 31.5% 90th percentile rate; 18.3% 75th percentile rate/upper quartile; 9.8% 50th percentile rate/median rate; 5% 25th percentile rate/lower quartile; and 1.5% 10th percentile rate.
CY 2012: 1,3404 denominator cases; 1,081 numerator cases; 170 hospitals; 8.0% national aggregate rate; 0.08296 mean of hospital rates; 0.09555 standard deviation; 21.2% 90th percentile rate; 10.8% 75th percentile rate/upper quartile; 4.9% 50th percentile rate/median rate; 2.6% 25th percentile rate/lower quartile; and 0% 10th percentile rate.
CY 2013: 1,4880 denominator cases; 658 numerator cases; 200 hospitals; 4.4% national aggregate rate; 0.05737 mean of hospital rates; 0.10193 standard deviation; 13.9% 90th percentile rate; 7.6%% 75th percentile rate/upper quartile; 2.6% 50th percentile rate/median rate; 0% 25th percentile rate/lower quartile; and 0% 10th percentile rate.
CY 2014: 130,882 denominator cases; 4,331 numerator cases; 1388 hospitals; 3.3% national aggregate rate; 0.03406 mean of hospital rates; 0.04647 standard deviation; 8.7% 90th percentile rate; 4.5% 75th percentile rate/upper quartile; 2.1% 50th percentile rate/median rate; 0% 25th percentile rate/lower quartile; and 0% 10th percentile rate.

1b.3. **If no or limited performance data on the measure as specified is reported in 1b2, then provide a summary of data from the literature that indicates opportunity for improvement or overall less than optimal performance on the specific focus of measurement.**

Not Applicable

1b.4. **Provide disparities data from the measure as specified (current and over time) by population group, e.g., by race/ethnicity, gender, age, insurance status, socioeconomic status, and/or disability.** *(This is required for endorsement maintenance. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities include.) This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*
In a review of two studies by Engle & Kominiarek published in 2008, it was determined that race and ethnicity have an impact on early term deliveries. When comparing Non-Hispanic White births with Hispanic and Black births, they found that the Non-Hispanic White births at 36 weeks accounted for the largest increase in elective cesarean deliveries from 1992 and 2002. This accounted for 3.1% to 3.9% of the total births reviewed in their study.

Interestingly, an overall increase was noted for all three groups. The reason for the increase has not been determined; however, factors speculated to account for the increase include socioeconomic status, access to health care and maternal demand for elective delivery. The rise in induction of labor is present for all racial groups with the highest increase in non-Hispanic whites. Bailey, et al. (2014) recently reviewed data for 638 rural women, recruited prenatally from three counties in rural southern Appalachia, who delivered electively at = 37 weeks. Those delivered electively at early term were 7.7 times more likely to be low birth weight, 4.4 times more likely to have a neonatal intensive care unit admission, and 2.5 times more likely to develop jaundice. Patients living furthest from the hospital were most likely to deliver electively at <39 weeks. Although rates of elective deliveries <39 weeks were no higher than national rates, adjusted odds ratios (aOR) of associated admission to a neonatal intensive care unit doubled (aOR 4.4 vs aOR 2.2).

1b.5. **If no or limited data on disparities from the measure as specified is reported in 1b4, then provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement. Include citations.**
• Bailey, B., McCook, J. & Chaires, C. (2014). Burden of elective early-term births in rural Appalachia. South Med J. 2014 Oct;107(10):624-9.
• Engle, W.A. & Kominiarek, M.A. (2008). Late preterm infants, early term infants, and timing of elective deliveries. Clin Perinatol. 35:325-41.

1c. High Priority (previously referred to as High Impact)
The measure addresses:

- a specific national health goal/priority identified by DHHS or the National Priorities Partnership convened by NQF; OR
- a demonstrated high-priority (high-impact) aspect of healthcare (e.g., affects large numbers of patients and/or has a substantial impact for a smaller population; leading cause of morbidity/mortality; high resource use (current and/or future); severity of illness; and severity of patient/societal consequences of poor quality).

1c.1. **Demonstrated high priority aspect of healthcare**
Frequently performed procedure, Patient/societal consequences of poor quality
1c.2. **If Other:**

1c.3. **Provide epidemiologic or resource use data that demonstrates the measure addresses a high priority aspect of healthcare. List citations in 1c.4.**
For almost 3 decades, the American College of Obstetricians and Gynecologists (ACOG) and the American Academy of Pediatrics (AAP) have had in place a standard requiring 39 completed weeks gestation prior to elective delivery, either vaginal or operative (ACOG, 1996). In 2009, ACOG published guidelines listing some of the acceptable medical indications for early induction of labor. Early term deliveries (between 37 and 38 weeks gestation) have increased dramatically from 1990 through 2006 and account for 17.5% of live births in the U.S. (Davidoff et al., 2006). A survey conducted in 2007 of almost 20,000 births in Hospital Corporation of America (HCA) hospitals throughout the U.S. carried out in conjunction with the March of Dimes at the request of ACOG, revealed that almost 1/3 of all babies delivered in the United States (US) are electively delivered earlier than the recommended 39 weeks of gestation (Clark et al., 2009). These procedures are conducted without documented evidence supporting medical indication for early delivery. This number of elective cesarean deliveries under 39 weeks gestation, without medical indication, represents 5% of all deliveries in the U.S., those deliveries violating ACOG/AAP guidelines (Clark et al., 2009). Such deliveries result in unnecessary costs to the U.S. health care system approaching $1 billion annually (Clark, et al., 2010).

Clark, et.al. (2009) found that most early elective deliveries are for convenience, and result in significant short term neonatal morbidity (neonatal intensive care unit admission rates of 13- 21%). Clark conducted a subsequent retrospective cohort study examining 27 hospitals, and determined that when strategies were implemented to reduce non-medically indicated elective early

term deliveries, there was a reduction in elective deliveries of 9.6% to 4.3% (Clark, et. al., 2010). Consequently, the rate of term neonatal intensive care admissions also fell by 16%.

According to Glantz (2005), when comparing spontaneous labor, elective inductions result in more cesarean deliveries and longer maternal length of stay. The American Academy of Family Physicians (2000) also notes that elective induction doubles the cesarean delivery rate. Repeat elective cesarean deliveries before 39 weeks gestation also results in higher rates of respiratory distress syndrome (RDS), mechanical ventilation, sepsis, and hypoglycemia for the newborns (Tita, et. al., 2009). Newborns born at 37 weeks gestation have a 7.5 fold greater rate of developing RDS versus those born at 39 to 41 weeks gestation (Tita, et al., 2009). Early-term newborns born at 37-38 weeks gestation also are at higher risk for transient tachypnea of the newborn, pulmonary hypertension, hospital stays greater than 5 days as well as diagnoses associated with severe morbidities or death versus newborns delivered at 39 weeks gestation (Engle & Kominiarek, 2008).

Recently, Little et al. (2015) conducted a multi-state analysis of data from 2005 through 2011 and MacDorman et al. (2015) analyzed U.S. data collected by the National Institutes of Health from 2006 through 2012. Both authors concluded that the lack of change in prospective stillbirth rates during these time periods suggests that preventing nonmedically indicated deliveries before 39 weeks of gestation did not increase the U.S. stillbirth rate.

1c.4. Citations for data demonstrating high priority provided in 1a.3

• American Academy of Family Physicians. (2000). Tips from Other Journals: Elective induction doubles cesarean delivery rate, 61, 4.Retrieved September 16, 2011 at: http://www.aafp.org/afp/20000215/tips/39.html.
• American Academy of Pediatrics, American College of Obstetricians and Gynecologists. Guidelines for perinatal care. 7th ed. Elk Grove Village (IL): AAP; Washington, DC: American College of Obstetricians and Gynecologists (2012). p. 109-110, 160, 192-194, 248.
• American College of Obstetricians and Gynecologists (ACOG) Practice Bulletin No 107. (2009). Induction of labor. Obstetrics & Gynecology. 114(2). 386-97.
• American College of Obstetricians and Gynecologists. (2013). Medically indicated late-preterm and early-term deliveries. Committee Opinion No. 560. Am J Obstet Gynecol 2013;121:908–10.
• American College of Obstetricians and Gynecologists. (2013). Nonmedically indicated early-term deliveries. Committee Opinion No. 561. Am J Obstet Gynecol. 121:911–5.
• Clark, S., Miller, D., Belfort, M., Dildy, G., Frye, D., & Meyers, J. (2009). Neonatal and maternal outcomes associated with elective delivery. [Electronic Version]. Am J Obstet Gynecol. 200:156.e1-156.e4.
• Clark, S., Frye, D., Meyers, J., Belfort, M., Dildy, G., Kofford, S et al. (2010). Reduction in elective delivery at <39 weeks of gestation: comparative effectiveness of 3 approaches to change and the impact on neonatal intensive care admission and stillbirth. Am J Obstet Gynecol. 203:449.e1-6.
• Clark, S., Meyers, J., Perlin, J. (2011). Oversight of elective early term deliveries – avoiding unintended consequences, Am J Obstet Gynecol, doi: 10.1016/j.ajog.2011.08.017.
• Clark, S., Meyers, J., Milton, C., Frye, D., Horner, S., Baker, A., & Perlin, J. (2013). Validation of the joint commission exclusion criteria for elective early-term delivery, Am J Obstet Gynecol.0:0. 1-5.
• Davidoff, M., Dias, T., Damus, K., Russell, R., Bettegowda, V.R., Dolan, S., et al. (2006). Changes in the gestational age distribution among U.S. singleton births: impacts on rates of late preterm birth, 1992-2002. Semin Perinatol. Feb; 30(1):8-15.
• Engle, W.A. & Kominiarek, M.A. (2008). Late preterm infants, early term infants, and timing of elective deliveries. Clin Perinatol. 35:325-41.
• Glantz, J. (Apr.2005). Elective induction vs. spontaneous labor associations and outcomes. [Electronic Version]. J Reprod Med. 50(4):235-40.
• Little, S., Zera, C., Clapp, M., Wilkins-Haug, L. & Robinson, J. (2015). A multi-state analysis of early-term delivery trends and the association with term stillbirth. Am J Obstet Gynecol. doi: 10.1097/AOG.0000000000001109.
• MacDorman, M., Reddy, U. & Silver, R. (2015). Trends in stillbirth by gestational age in the united states, 2006-2012. Am J Obstet Gynecol. doi: 10.1097/AOG.0000000000001152.
• Tita, A., Landon, M., Spong, C., Lai, Y., Leveno, K., Varner, M, et al. (2009). Timing of elective repeat cesarean delivery at term and neonatal outcomes. [Electronic Version]. NEJM. 360:2, 111-120.

1c.5. If a PRO-PM (e.g. HRQoL/functional status, symptom/burden, experience with care, health-related behaviors), provide evidence that the target population values the measured PRO and finds it meaningful. *(Describe how and from whom their input was obtained.)*

Not Applicable

2. Reliability and Validity—Scientific Acceptability of Measure Properties

Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. ***Measures must be judged to meet the subcriteria for both reliability and validity to pass this criterion and be evaluated against the remaining criteria.***

2a.1. Specifications The measure is well defined and precisely specified so it can be implemented consistently within and across organizations and allows for comparability. eMeasures should be specified in the Health Quality Measures Format (HQMF) and the Quality Data Model (QDM).

De.5. **Subject/Topic Area** *(check all the areas that apply):*
Perinatal and Reproductive Health, Perinatal and Reproductive Health : Perinatal

De.6. **Cross Cutting Areas** *(check all the areas that apply):*
Overuse, Prevention : Social Determinants, Safety : Complications

S.1. **Measure-specific Web Page** *(Provide a URL link to a web page specific for this measure that contains current detailed specifications including code lists, risk model details, and supplemental materials. Do not enter a URL linking to a home page or to general information.)*
http://manual.jointcommission.org/releases/TJC2015B2/

S.2a. **If this is an eMeasure**, HQMF specifications must be attached. Attach the zipped output from the eMeasure authoring tool (MAT) - if the MAT was not used, contact staff. (Use the specification fields in this online form for the plain-language description of the specifications)
This is not an eMeasure **Attachment:**

S.2b. **Data Dictionary, Code Table, or Value Sets** *(and risk model codes and coefficients when applicable) must be attached. (Excel or csv file in the suggested format preferred - if not, contact staff)*
Attachment **Attachment:** PC01_ICD_Code_Tables.xlsx

S.3. **For endorsement maintenance**, please briefly describe any changes to the measure specifications since last endorsement date and explain the reasons.

- All ICD-9-CM diagnosis codes and ICD-9-CM procedure codes were converted to ICD-10-CM diagnosis and ICD-10-PCS procedure codes throughout the measure specifications. The Department of Health and Human Services (HHS) mandated that all entities covered by the Health Insurance Portability and Accountability Act (HIPAA) must transition to a new set of codes for electronic health care transactions on October 1, 2015.
- The initial patient population is now identified with ICD-10-PCS-Principal or Other Procedure Codes for delivery instead of diagnosis codes for pregnancy, since the ICD-10-CM Principal or Other Diagnosis Codes do not indicate whether the delivery took place during the hospitalization.
- The numerator included population for medical induction of labor now requires a check for the presence of labor prior to the procedure, since the ICD-10-PCS-Principal and Other Procedure Codes assigned for labor induction do not distinguish between induction and augmentation of labor.
- The numerator data element Spontaneous Rupture of Membranes was removed as a check prior to medical induction of labor, since this data element was duplicative of existing ICD-10-CM Diagnosis Codes in Appendix on Table 11.07 for premature and prolonged rupture of membranes which should be applied to exclude such cases from the measure.
- A new numerator data element Prior Uterine Surgery was added for cases with cesarean births in order to pass cases with acceptable prior uterine surgeries as noted below under S.6 numerator details.
- Cases with a gestational age of UTD were added to the denominator excluded populations, since UTD is highly correlated with no prenatal care.
- New ICD-10-CM Diagnosis Codes were added to Table 11.07: Conditions Possibly Justifying Elective Delivery Prior to 39 Weeks Gestation in Appendix A based on feedback from the field and recommendations from the PC Technical Advisory Panel.
- Vital Records reports, delivery logs and clinical information systems were added as acceptable data sources to help hospitals identify all cases with 37 and 38 weeks gestation, so that 100% of these cases could be reviewed to increase the denominator population size.

S.4. **Numerator Statement** *(Brief, narrative description of the measure focus or what is being measured about the target population, i.e., cases from the target population with the target process, condition, event, or outcome)*
IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the

calculation algorithm.
Patients with elective deliveries with ICD-10-PCS Principal Procedure Code or ICD-10-PCS Other Procedure Codes for one or more of the following:
• Medical induction of labor as defined in Appendix A, Table 11.05 available at:
http://manual.jointcommission.org/releases/TJC2015B2/ while not in Labor prior to the procedure
• Cesarean birth as defined in Appendix A, Table 11.06 and all of the following:
not in Labor
no history of a Prior Uterine Surgery available at: http://manual.jointcommission.org/releases/TJC2015B2/

S.5. **Time Period for Data** *(What is the time period in which data will be aggregated for the measure, e.g., 12 mo, 3 years, look back to August for flu vaccination? Note if there are different time periods for the numerator and denominator.)*
Episode of care

S.6. **Numerator Details** *(All information required to identify and calculate the cases from the target population with the target process, condition, event, or outcome such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
IF an OUTCOME MEASURE, describe how the observed outcome is identified/counted. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
Four data elements are used to calculate the numerator:

1. ICD-10-PCS Other Procedure Codes - The International Classification of Diseases, Tenth Revision, Procedure Coding System code that identifies significant procedures performed other than the principal procedure during this hospitalization.
2. ICD-10-PCS Principal Procedure Code - The International Classification of Diseases, Tenth Revision, Procedure Coding System code that identifies the principal procedure performed during this hospitalization. The principal procedure is the procedure performed for definitive treatment rather than diagnostic or exploratory purposes, or which is necessary to take care of a complication.
3. Labor- Documentation that the patient was in labor prior to induction and/or cesarean birth. Allowable values: Yes or No/UTD.
4. Prior Uterine Surgery- Documentation that the patient had undergone prior uterine surgery which includes: a prior classical cesarean birth defined as a vertical incision into the upper uterine segment, a prior myomectomy, a prior uterine surgery resulting in a perforation of the uterus due to an accidental injury, a history of a uterine window or thinning of the uterine wall noted during prior uterine surgery or during ultrasound, a history of uterine rupture requiring surgical repair, a history of a cornual ectopic pregnancy or history of a transabdominal cerclage.
Allowable Values: Yes or No/UTD

Patients are eligible for the numerator population with ICD-10-PCS Other Procedure Codes or ICD-10-PCS Principal Procedure Code for medical induction or with ICD-10-PCS Other Procedure Codes or ICD-10-PCS Principal Procedure Code for cesarean birth when the allowable value equals “no” for the data elements Labor and Prior Uterine Surgery.
Updates available at: http://manual.jointcommission.org/releases/TJC2015B2/

S.7. **Denominator Statement** *(Brief, narrative description of the target population being measured)*
Patients delivering newborns with >= 37 and < 39 weeks of gestation completed with ICD-10-PCS Principal or Other Procedure Codes for delivery as defined in Appendix A, Table 11.01.1 available at: http://manual.jointcommission.org/releases/TJC2015B2/ and with ICD-10-CM Principal Diagnosis Code or ICD-10-CM Other Diagnosis Codes for planned cesarean birth in labor as defined in Appendix A, Table 11.06.1 available at: http://manual.jointcommission.org/releases/TJC2015B2/

S.8. **Target Population Category** *(Check all the populations for which the measure is specified and tested if any):*
Maternal Health

S.9. **Denominator Details** *(All information required to identify and calculate the target population/denominator such as definitions, specific data collection items/responses , code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
Seven data elements are used to calculate the denominator:

1. Admission Date – The month, day and year of admission to acute inpatient care.
2. Birthdate - The month, day and year the patient was born.
3. Clinical Trial - Documentation that during this hospital stay the patient was enrolled in a clinical trial in which patients with pregnancy were being studied. Allowable values: Yes or No/UTD

4. Discharge Date – The month day and year the patient was discharged from acute care, left against medical advice or expired during the stay.
5. Gestational Age – Documentation of the weeks of gestation completed at the time of delivery. Allowable Values: 1-50 or UTD.
6. ICD-10-CM Other Diagnosis Codes - The International Classification of Diseases, Tenth Revision, Clinical Modification codes associated with the secondary diagnoses for this hospitalization.
7. ICD-10-CM Principal Diagnosis Code - The International Classification of Diseases, Tenth Revision, Clinical Modification code associated with the diagnosis established after study to be chiefly responsible for occasioning the admission of the patient for this hospitalization.
Updates available at: http://manual.jointcommission.org/releases/TJC2015B2/

S.10. **Denominator Exclusions** *(Brief narrative description of exclusions from the target population)*
• ICD-10-CM Principal Diagnosis Code or ICD-10-CM Other Diagnosis Codes for conditions possibly justifying elective delivery prior to 39 weeks gestation as defined in Appendix A, Table 11.07
• Less than 8 years of age
• Greater than or equal to 65 years of age
• Length of Stay >120 days
• Enrolled in clinical trials
• Gestational Age < 37 or >= 39 weeks or UTD

S.11. **Denominator Exclusion Details** *(All information required to identify and calculate exclusions from the denominator such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
• Patients with ICD-10-CM Principal Diagnosis Code or Other Diagnosis Codes for conditions for possibly justifying elective delivery are excluded.
• The patient age in years is equal to the Admission Date minus the Birthdate. Patients less than 8 years of age or greater or equal to 65 years of age are excluded.
• Length of stay (LOS) in days is equal to the Discharge Date minus the Admission Date. If the LOS is greater than 120 days, the patient is excluded.
• Patients are excluded if "Yes" is selected for Clinical Trial.
• Patients with a Gestational Age less than 37 weeks or equal to or greater than 39 weeks or UTD are excluded from the measure.

S.12. **Stratification Details/Variables** *(All information required to stratify the measure results including the stratification variables, definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format with at S.2b)*
Not Applicable

S.13. **Risk Adjustment Type** (Select type. Provide specifications for risk stratification in S.12 and for statistical model in S.14-15)
No risk adjustment or risk stratification
If other:

S.14. **Identify the statistical risk model method and variables** *(Name the statistical method - e.g., logistic regression and list all the risk factor variables. Note - risk model development and testing should be addressed with measure testing under Scientific Acceptability)*
Not Applicable

S.15. **Detailed risk model specifications** *(must be in attached data dictionary/code list Excel or csv file. Also indicate if available at measure-specific URL identified in S.1.)*
Note: Risk model details (including coefficients, equations, codes with descriptors, definitions), should be provided on a separate worksheet in the suggested format in the Excel or csv file with data dictionary/code lists at S.2b.

S.15a. **Detailed risk model specifications** (*if not provided in excel or csv file at S.2b*)

S.16. **Type of score:**
Rate/proportion
If other:

S.17. **Interpretation of Score** *(Classifies interpretation of score according to whether better quality is associated with a higher score, a lower score, a score falling within a defined interval, or a passing score)*
Better quality = Lower score

S.18. **Calculation Algorithm/Measure Logic** *(Describe the calculation of the measure score as an ordered sequence of steps including identifying the target population; exclusions; cases meeting the target process, condition, event, or outcome; aggregating data; risk adjustment; etc.)*
1. Start processing. Run cases that are included in the PC-Mother Initial Patient Population and pass the edits defined in the Transmission Data Processing Flow: Clinical through this measure.

2. Check ICD-10-CM Principal or Other Diagnosis Codes
a. If at least one of the ICD-10-CM Principal or Other Diagnosis Codes is on Table 11.07, the case will proceed to a Measure Category Assignment of B and will not be in the Measure Population. Stop processing.
b. If none of the ICD-10-CM Principal or Other Diagnosis Code is on Table 11.07, continue processing and proceed to Clinical Trial.

3. Check Clinical Trial
a. If Clinical Trial is missing, the case will proceed to a Measure Category Assignment of X and will be rejected. Stop Processing.
b. If Clinical Trial equals Yes, the case will proceed to a Measure Category Assignment of B and will not be in the measure population. Stop Processing.
c. If Clinical Trial equals No, continue processing and proceed to Gestational Age.

4. Check Gestational Age
a. If Gestational Age is missing, the case will proceed to a Measure Category Assignment of X and will be rejected. Stop Processing.
b. If Gestational Age is less than 37 or greater than or equal to 39 or equal to a Not Unable to Determine Value, the case will proceed to a Measure Category Assignment of B and will not be in the measure population. Stop Processing.
c. If Gestational Age is greater than or equal to 37 and less than 39, continue processing and proceed to recheck ICD-10-CM Principal Procedure or Other Diagnosis Codes.

5. Recheck ICD-10-CM Principal or Other Diagnosis Codes
a. If at least one of the ICD-10-CM Principal or Other Diagnosis Code is on Table 11.06.1, the case will proceed to a Measure Category Assignment of D and will
be in the Measure Population. Stop processing.
b. If none of the ICD-10-CM Principal or Other Diagnosis Code is on Table 11.06.1, continue processing and proceed to ICD-10-CM Principal or Other Procedure Codes.

6. Check ICD-10-PCS Principal or Other Procedure Codes
a. If all of the ICD-10-PCS Principal or Other Procedure Codes are missing, the case will proceed to a Measure Category Assignment of D and will be in the Measure Population. Stop Processing.
b. If at least one of the ICD-10-PCS Principal or Other Procedure Codes is on Table 11.05, continue processing and proceed to Labor
i. If Labor is missing, the case will proceed to a Measure Category Assignment of X and will be rejected. Stop Processing.
ii. If Clinical Trial equals Yes, the case will proceed to a Measure Category Assignment of D and will be in the Measure Population. Stop Processing.
iii. If Labor equals No, the case will proceed to a Measure Category Assignment of E and will be in the Numerator Population. Stop Processing.
c. If none of the ICD-9-CM Principal Procedure Codes is on Table 11.05, continue processing and proceed to recheck ICD-10-PCS Principal or Other Procedure Codes.

7. Recheck ICD-10-PCS Principal or Other Procedure Codes
a. If none of the ICD-10-PCS Principal or Other Procedure Code is on Table 11.06, the case will proceed to a Measure Category Assignment of D and will be in the Measure Population. Stop Processing.
b. If at least one of the ICD-10-PCS Principal or Other Procedure Code is on Table 11.06, continue processing and proceed to Labor.

8. Check Labor
a. If Labor is missing, the case will proceed to a Measure Category Assignment of X and will be rejected. Stop Processing.
b. If Labor equals Yes, the case will proceed to a Measure Category Assignment of D and will be in the Measure Population. Stop processing.
c. If Labor equals No, continue processing and proceed to Spontaneous Rupture of Membranes.

9. Check Prior Uterine Surgery
a. If Prior Uterine Surgery is missing, the case will proceed to a Measure Category Assignment of X and will be rejected. Stop Processing.
b. If Prior Uterine Surgery equals Yes, the case will proceed to a Measure Category Assignment of D and will be in the Measure Population. Stop processing.
c. If Prior Uterine Surgery equals No, the case will proceed to a Measure Category Assignment of E and will be in the Numerator Population. Stop Processing.

S.19. **Calculation Algorithm/Measure Logic Diagram URL or Attachment** *(You also may provide a diagram of the Calculation Algorithm/Measure Logic described above at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
Available at measure-specific web page URL identified in S.1

S.20. **Sampling** *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*
IF a PRO-PM, identify whether (and how) proxy responses are allowed.
The initial patient population includes patients admitted to the hospital for inpatient acute care if they have: ICD-10-PCS Principal or Other Procedure Code as defined in Appendix A, Table 11.01.1, a Patient Age (Admission Date – Birthdate) >= 8 years and < 65 and a Length of Stay (Discharge Date - Admission Date) = 120 days. The sample is taken randomly as follows for a monthly sample:
• Average monthly Initial Patient Population >= 501 results in a minimum random sample size of 101.
• Average monthly Initial Patient Population 126 – 500 results in a minimum random sample size of 20% of the population size.
• Average monthly Initial Patient Population 25 – 125 results in a minimum random sample size of 25.
• Average monthly Initial Patient Population < 25 results in no sampling; 100% Initial Patient Population required

S.21. **Survey/Patient-reported data** *(If measure is based on a survey, provide instructions for conducting the survey and guidance on minimum response rate.)*
IF a PRO-PM, specify calculation of response rates to be reported with performance measure results.
Not Applicable

S.22. **Missing data** (specify how missing data are handled, e.g., imputation, delete case.)
Required for Composites and PRO-PMs.
Any file with missing data will result in a measure category assignment of X and rejection of the file from the warehouse, unless the data are not used to process the measure. If the data are used to process the measure and have been reported by the abstractor as No/UTD, the case will result in a measure category assignment of E (i.e., failed measure, not rejected).

S.23. **Data Source** *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*
If other, please describe in S.24.
Electronic Clinical Data, Paper Medical Records

S.24. **Data Source or Collection Instrument** *(Identify the specific data source/data collection instrument e.g. name of database, clinical registry, collection instrument, etc.)*
IF a PRO-PM, identify the specific PROM(s); and standard methods, modes, and languages of administration.
Each data element in the data dictionary includes suggested data sources. The data are collected using contracted Performance Measurement Systems (vendors) that develop data collection tools based on the measure specifications. The tools are verified and tested by Joint Commission staff to confirm the accuracy and conformance of the data collection tool with the measure specifications. The vendor may not offer the measure set to hospitals until verification has been passed.

S.25. **Data Source or Collection Instrument** *(available at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
No data collection instrument provided

S.26. **Level of Analysis** *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*
Facility, Population : National

S.27. **Care Setting** *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*
Hospital/Acute Care Facility
If other:

S.28. **COMPOSITE Performance Measure** - Additional Specifications *(Use this section as needed for aggregation and weighting rules, or calculation of individual performance measures if not individually endorsed.)*

Not Applicable

2. RELIABILITY & VALIDITY - SCIENTIFIC ACCEPTABILITY OF MEASURE PROPERTIES

Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. **(evaluation criteria)**
Measure testing must demonstrate adequate reliability and validity in order to be recommended for endorsement. Testing may be conducted for data elements and/or the computed measure score. Testing information and results should be entered in the appropriate field. Supplemental materials may be referenced or attached in item 2.1. See guidance on measure testing.

2a2. Reliability Testing. (*Reliability testing was conducted with appropriate method, scope, and adequate demonstration of reliability.*)

2a2.1 Data/Sample *(Description of the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)***:**
The PC measure set has been in national use since the 2nd quarter of 2010. It is a requirement of participation in the ORYX initiative that data on all measures in the set are collected. (ORYX is the term used by The Joint Commission to describe the component of the hospital accreditation program which requires data collection and reporting on standardized national performance measures.)
Demographics of organizations collecting and reporting data on these measures are as follows:
163 health care organizations representing various types, locations and sizes:
10 For Profit, 91 Not for Profit, 46 Military Facilities, 9 County, 2 State, 5 Other
15 >=500 beds; 29 250-499 beds; 50 100-249 beds; 69 <100 beds
Located in: AE, AK, AL, AP, AR, AZ, CA, DO, DC, FL, GA, HI, ID, IL, IN, KS, KY, LA, MA, MD, MI, MN, MO, MS, MT, NC, NE, NV, NY, OH,OK, PA, PR, RI, SC, TN, TX, VA, WA, WI, WV
26 performance measurement systems

2a2.2 Analytic Method *(Describe method of reliability testing & rationale)***:**
This measure was adapted from NQF-endorsed measure 0469 Elective Delivery Prior to 39 Completed Weeks Gestation. As such, initial data reliability would have been addressed during the original endorsement. The Joint Commission will be conducting additional reliability studies on this measure as well as the entire PC measure set beginning in October 2011.
Currently, hospitals are supported in their data collection and reporting efforts by 26 contracted performance measurement system (PMS) vendors. It is a contractual requirement of Joint Commission listed vendors that the quality and reliability of data submitted to them by contracted health care organizations must be monitored on a quarterly basis. In addition, The Joint Commission analyzes these data by running 17 quality tests on the data submitted into ORYX. (ORYX is the term used by The Joint Commission to describe the component of the hospital accreditation program which requires data collection and reporting on standardized national performance measures). The following is a list of the major tests done on the submitted ORYX data, taken from the 2011 ORYX Performance Measurement System Requirements manual.

- Transmission of complete data
- Usage of individual core measure data received: To understand if the HCO provides the relevant service to treat the relevant population
- Investigation of aberrant data points
- Verification of patient population and sample size
- Identification of missing data elements
- Validation of the accuracy of target outliers
- Data integrity
- Data corrections

Data Element Agreement Rate:
Inter-rater reliability testing methodology utilized by contracted performance measure system vendors as outlined in the contract is as follows:
- All clinical data elements and all editable demographic elements are scored.
- All measure data are reabstracted with originally abstracted data having been blinded so that the reabstraction is not biased.
- Reabstracted data are compared with originally abstracted data on a data element by data element basis. A data element agreement rate is calculated. Clinical and demographic data are scored separately, and an overall agreement rate is computed.

2a2.3 Testing Results *(Reliability statistics, assessment of adequacy in the context of norms for the test conducted)***:**
Data element agreement rates were reported to The Joint Commission for 1Q11 This reflects the findings of 108 hospitals, comprising 13,279 records (100% sample). The following table delineates calculated agreement rates for individual data elements that are used to compute measure rates for PC-01.
Data Elements with a Mismatch - Mother total n total d rate

Active Labor	33 35	94.29%	
Gestational Age	639	712	89.75%

These agreement rates are considered to be well within acceptable levels.

2b. VALIDITY. Validity, Testing, including all Threats to Validity: H☐ M☐ L☐ I ☐

2b1.1 Describe how the measure specifications *(measure focus, target population, and exclusions)* **are consistent with the evidence cited in support of the measure focus (***criterion 1c)* **and identify any differences from the evidence:**
This measure focuses on reducing non-medically indicated elective deliveries at >= 37 and < 39 weeks of gestation completed. The literature supports the focus on patients delivering newborns within this gestational age range. Accordingly, this measure excludes patients with conditions possibly justifying elective delivery. Also excluded from the measure are patients with a length of stay greater than 120 days, and those enrolled in a clinical trial. These exclusions are not addressed in the literature, but are included for this measure in order to harmonize with other CMS/Joint Commission aligned measures.

2b2. Validity Testing. (*Validity testing was conducted with appropriate method, scope, and adequate demonstration of validity.)*

See below updated valildity testing

2b2.1 Data/Sample *(Description of the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)***:**
As previously mentioned the PC measure set has been in national use since the 2nd quarter of 2010. Demographics of organizations collecting and reporting data on these measures are as follows:
163 health care organizations representing various types, locations and sizes:
10 For Profit, 91 Not for Profit, 46 Military Facilities, 9 County, 2 State, 5 Other
15 >=500 beds; 29 250-499 beds; 50 100-249 beds; 69 <100 beds

Located in: AE, AK, AL, AP, AR, AZ, CA, DO, DC, FL, GA, HI, ID, IL, IN, KS, KY, LA, MA, MD, MI, MN, MO, MS, MT, NC, NE, NV, NY, OH,OK, PA, PR, RI, SC, TN, TX, VA, WA, WI, WV
26 performance measurement systems

2b2.2 Analytic Method *(Describe method of validity testing and rationale; if face validity, describe systematic assessment)***:**
Since the measure has been in national use, continued face validity of the measure has been determined through analysis of feedback from measure users. The Joint Commission provides a web-based application with which measure users can provide feedback regarding appropriateness of measure specifications, request clarification of specifications, and/or provide other comments pertinent to the measure. This feedback is systematically, continually, reviewed in order to identify trends and to identify areas of the measure specifications that require clarification or revision. Additionally, Joint Commission staff continually monitors the national literature and environment in order to assess continued validity of this measure.

As noted previously, The Joint Commission is currently performing reliability site visits. A component of these visits will include focus group interviews with hospital staff working with the PC measures to obtain feedback regarding the validity of the measures and suggestions for further refinement of the specifications.

2b2.3 Testing Results *(Statistical results, assessment of adequacy in the context of norms for the test conducted; if face validity, describe results of systematic assessment)***:**
Analysis of feedback obtained via our automated feedback system reveals slightly more than 180 submissions regarding specifications for this measure since its implementation in 2010. Predominant themes of these submissions involved questions regarding clarification of the data elements Active Labor and Gestational Age with respect to both definitions and the calculation of gestational age and the order of priority sources to retrieve the data. Additional notes for abstractors were added to the data elements for clarification. In addition, the data elements Active Labor and Spontaneous Rupture of Membranes were moved from the numerator population to the denominator population and the algorithm was revised in order to capture all deliveries in the denominator population. Additional ICD-9-CM diagnosis codes were added to Table 11.07 to update exclusions based on consultation with the perinatal care experts. The gestational age range for the denominator statement and included population was also revised to exclude patients with a gestational age of 39 weeks of gestation completed, since the upper range for gestational age for 38 weeks ends at 38 6/7 weeks gestation.

POTENTIAL THREATS TO VALIDITY. (*All potential threats to validity were appropriately tested with adequate results.*)

2b3. Measure Exclusions. (*Exclusions were supported by the clinical evidence in 1c or appropriately tested with results*

demonstrating the need to specify them.)

2b3.1 Data/Sample for analysis of exclusions *(Description of the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)***:**
As previously mentioned the PC measure set has been in national use since the 2nd quarter of 2010. Demographics of organizations collecting and reporting data on these measures are as follows:
163 health care organizations representing various types, locations and sizes:
10 For Profit, 91 Not for Profit, 46 Military Facilities, 9 County, 2 State, 5 Other
15 >=500 beds; 29 250-499 beds; 50 100-249 beds; 69 <100 beds
Located in: AE, AK, AL, AP, AR, AZ, CA, DO, DC, FL, GA, HI, ID, IL, IN, KS, KY, LA, MA, MD, MI, MN, MO, MS, MT, NC, NE, NV, NY, OH,OK, PA, PR, RI, SC, TN, TX, VA, WA, WI, WV
26 performance measurement systems

2b3.2 Analytic Method (*Describe type of analysis and rationale for examining exclusions, including exclusion related to patient preference)***:**
Measure exclusions that were not derived directly from the evidence are presented below. Please note that these are population exclusions that are necessary to ensure consistency in all measures in this 5 measure set.
These exclusions were analyzed for frequency of occurrence. An issue that is of great concern to users of this measure is that due to the presence of exceptions to the measure, attainment of a 0% measure rate is not possible. Because of the role of this measure in the current Joint Commission accreditation process this is especially troubling to measure users. This concern is the basis for a number of the non-evidence-based exclusions to these measures. Additional reasons for these population exclusions are enumerated in our response to section 2b1.1 above. The following measure exclusions that were not derived directly from the evidence are as follows:

1. Patients with LOS <120 days
2. Patients less than 8 years of age or greater than or equal to 65 years of age
3. Patients enrolled in clinical trials

2b3.3 Results *(Provide statistical results for analysis of exclusions, e.g., frequency, variability, sensitivity analyses):*
N=353,671
1. Patients who have a length of stay (LOS) greater than 120 days =0%
2. Patients less than 8 years of age or greater than or equal to 65 years of age=0%
3. Patients enrolled in clinical trials =0.04%

2b4. Risk Adjustment Strategy. (*For outcome measures, adjustment for differences in case mix (severity) across measured entities was appropriately tested with adequate results.*)

2b4.1 Data/Sample *(Description of the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)***:**
Not Applicable

2b4.2 Analytic Method (*Describe methods and rationale for development and testing of risk model or risk stratification including selection of factors/variables)***:**
Not Applicable

2b4.3 Testing Results *(Statistical risk model: Provide quantitative assessment of relative contribution of model risk factors; risk model performance metrics including cross-validation discrimination and calibration statistics, calibration curve and risk decile plot, and assessment of adequacy in the context of norms for risk models. Risk stratification: Provide quantitative assessment of relationship of risk factors to the outcome and differences in outcomes among the strata)***:**
Not Applicable

2b4.4 If outcome or resource use measure is not risk adjusted, provide rationale and analyses to justify lack of adjustment:
Not Applicable

2b5. Identification of Meaningful Differences in Performance. (*The performance measure scores were appropriately analyzed and discriminated meaningful differences in quality.*)

2b5.1 Data/Sample *(Describe the data or sample including number of measured entities; number of patients; dates of data; if a*

sample, characteristics of the entities included):
As previously mentioned the PC measure set has been in national use since the 2nd quarter of 2010. Demographics of organizations collecting and reporting data on these measures are as follows:
163 health care organizations representing various types, locations and sizes:
10 For Profit, 91 Not for Profit, 46 Military Facilities, 9 County, 2 State, 5 Other
15 >=500 beds; 29 250-499 beds; 50 100-249 beds; 69 <100 beds
Located in: AE, AK, AL, AP, AR, AZ, CA, DO, DC, FL, GA, HI, ID, IL, IN, KS, KY, LA, MA, MD, MI, MN, MO, MS, MT, NC, NE, NV, NY, OH,OK, PA, PR, RI, SC, TN, TX, VA, WA, WI, WV
26 performance measurement systems

2b5.2 Analytic Method *(Describe methods and rationale to identify statistically significant and practically/meaningfully differences in performance)*:
The method used to analyze meaningful differences in performance at The Joint Commission is Target Analysis. The object of target analysis is to compare a health care organization's (HCO) data against a comparative norm for the purpose of evaluating performance improvement opportunities. When an organization's performance level is statistically significantly different from a comparative norm, it is considered a statistical deviation. A statistical deviation may be desirable or undesirable depending on the "direction of improvement" of the measure.
There are two components to the target analysis methodology used at The Joint Commission. Given the national average for a performance measure, a target range is constructed. Using generalized linear mixed models methodology (also known as hierarchical models), a predicted estimate of an HCO's performance, with a corresponding 95% confidence interval, is generated. This confidence interval is compared to the target range, to determine the HCOs' rating. The estimate of the organization's true performance is based on both the data from that organization and on data from the entire set of reporting organizations.

2b5.3 Results *(Provide measure performance results/scores, e.g., distribution by quartile, mean, median, SD, etc.; identification of statistically significant and meaningfully differences in performance)*:
PC-01 Distribution of Outliers

2011 1st Quarter Data:
Scores on this measure: N=160, Mean 13.6%, SD 0.1594
10th Percentile= 0%
25th Percentile= 0%
50th Percentile= 9%
75th Percentile= 19%
90th Percentile= 34%

156 (97.5%) Neutral – results not significantly different from target range
4 (2.5%) Unfavorable - results statistically significantly lower than the national rate

2b6. Comparability of Multiple Data Sources/Methods. (*If specified for more than one data source, the various approaches result in comparable scores.*)

2b6.1 Data/Sample *(Describe the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)*:
Multiple data sources are not used for this measure.

2b6.2 Analytic Method *(Describe methods and rationale for testing comparability of scores produced by the different data sources specified in the measure)*:
Not Applicable

2b6.3 Testing Results *(Provide statistical results, e.g., correlation statistics, comparison of rankings; assessment of adequacy in the context of norms for the test conducted)*:
Not Applicable

2c. Disparities in Care: H☐ M☐ L☐ I ☐ NA☐ (*If applicable, the measure specifications allow identification of disparities.*)

2c.1 If measure is stratified for disparities, provide stratified results *(Scores by stratified categories/cohorts)*: The measure is not stratified for disparities.

2c.2 **If disparities have been reported/identified (e.g., in 1b), but measure is not specified to detect disparities, please explain:** Although a rise in induction of labor is present for all racial groups with the highest increase in non-Hispanic whites, there are no plans to stratify the measure. The Joint Commission does not currently capture date elements for race or ethnicity because these data elements have not been shown to be reliably collectable due to the fact that no national standardized definitions exist for these data elements. Also, not all hospitals collect race and ethnicity. In the future, it may be feasible for The Joint Commission to explore how race and ethnicity and other relevant disparity data, might be collected reliably in the future.
2.1-2.3 **Supplemental Testing Methodology Information:**
Steering Committee: Overall, was the criterion, *Scientific Acceptability of Measure Properties*, met? **(*Reliability and Validity must be rated moderate or high*) Yes☐ No☐** **Provide rationale based on specific subcriteria:**
If the Committee votes No, STOP

NEW (since 2012 endorsement): Data for Empirical Testing

2b2. VALIDITY TESTING

2b2.1. What level of validity testing was conducted? (*may be one or both levels*)

☐ **Critical data elements** (*data element validity must address ALL critical data elements*)

☐ **Performance measure score**

☒ **Empirical validity testing**

☐ **Systematic assessment of face validity of performance measure score as an indicator** of quality or resource use (*i.e., is an accurate reflection of performance on quality or resource use and can distinguish good from poor performance*)

2b2.2. For each level of testing checked above, describe the method of validity testing and what it tests (*describe the steps—do not just name a method; what was tested, e.g., accuracy of data elements compared to authoritative source, relationship to another measure as expected; what statistical analysis was used)*

ICD-9 to ICD-10 Conversion Process:
The goal was to convert ICD-9 to ICD-10 equivalent codes, consistent with the clinical intent of the original measure specifications. The Joint Commission worked with a certified coding expert throughout the conversion process. The 3M Coding Conversion Tool was utilized, including forward mapping of ICD-9 codes to ICD-10 codes as well as reverse mapping from ICD-10 to ICD-9 to ensure appropriateness. MSDRGs and instructions in the tabular index were also examined to ensure appropriate code mapping. Crosswalks comprising ICD-9 codes mapped to ICD-10 codes were created and reviewed by members of the Technical Advisory Panel, CMS subcontractors, and performance measurement system vendors prior to being posted for a 12 month public comment period. Feedback from the field indicated that the crosswalks generally were mapped correctly. Minor modifications to the code tables were made as needed. Final code tables were published in early 2015, well in advance of the mandated date of October 1, 2015.

Perinatal Care (PC) Initial Patient Population

The PC measure set is unique in that there are two distinct Initial Patient Populations within the measure set, mothers (PC-01, PC-02, PC-03) and newborns. (PC-04, PC-05).

Subpopulation Mothers

Patients admitted to the hospital for inpatient acute care are included in the PC Mother Initial sampling group if they have: ICD-9-CM Principal or Other Diagnosis Code as defined in Appendix A, Tables 11.01, 11.02, 11.03, or 11.04, a Patient Age (Admission Date Birthdate) >= 8 years and < 65 and a Length of Stay (Discharge Date - Admission Date) ≤ 120 days. PC-01- Elective Delivery belongs to the above population.

The data used to measure the validity of the PC measure are comprised of data from the third and fourth quarters of 2014, and the first and second quarters of 2015. 1,345 hospitals submitted 2,695,467 inpatient records for all the elected PC measures. The hospitals included in the analysis reported one year of data and had 30 or more denominator cases in the analysis period.

Measure convergent validity for PC-01 was assessed using hospitals patient level data from The Joint Commission warehouse. Measure specifications, including population identification, numerator and denominator statements, exclusions, and data elements and their definitions were found to be understandable, retrievable, and relevant in previous validity testing.

2b2.3. What were the statistical results from validity testing? (*e.g., correlation; t-test*)
Overall descriptive statistics for sub population MOTHER:

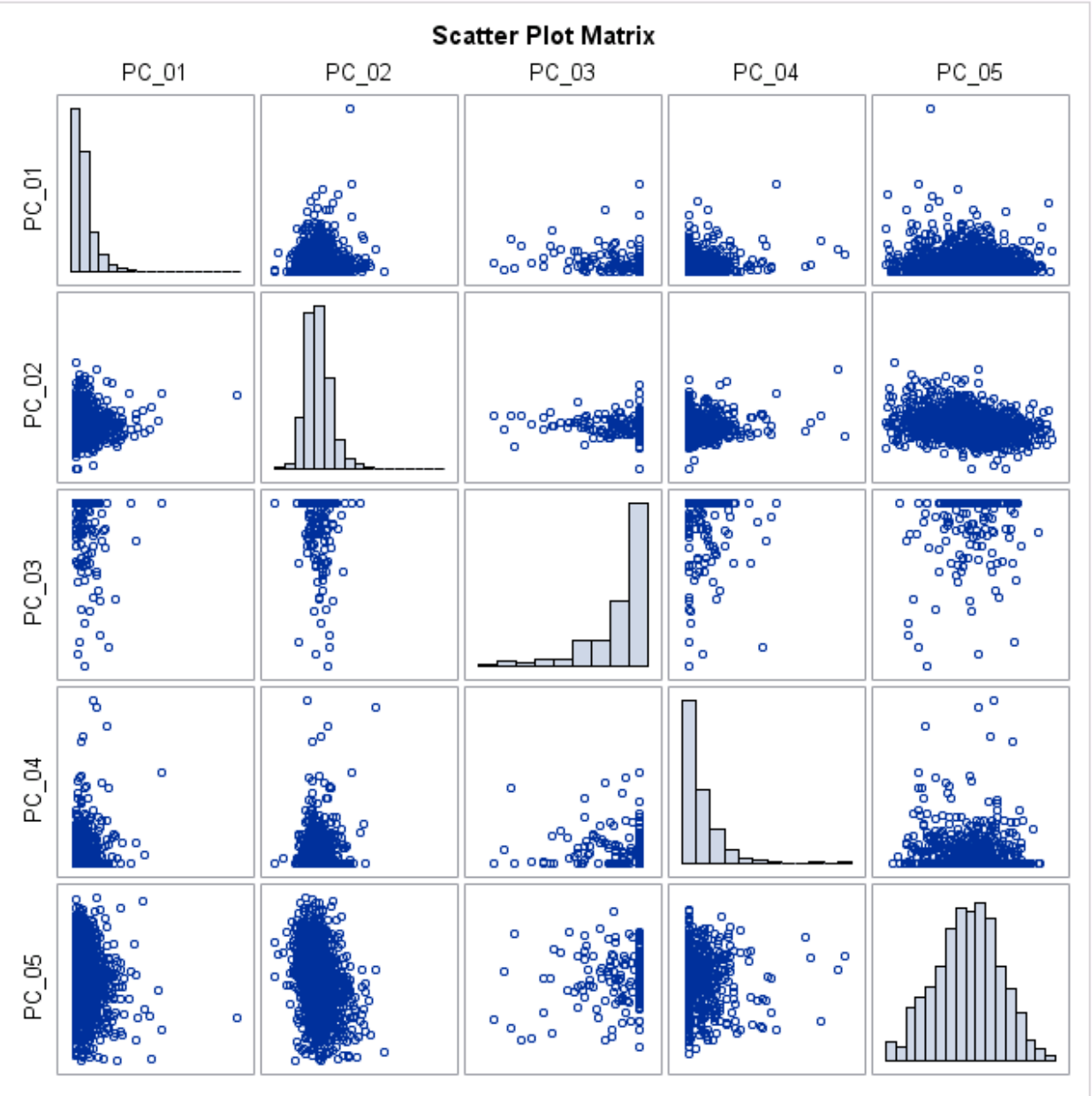
N=1,345 hospitals
n = 2,695,467 records submitted

Descriptive statistics for PC- 01 measure:
N = 1,237 hospitals
n = 1,130,083
Mean: 2.74%
Min = 0%
Percentile 10%: 0%
Percentile 25%: 0%
Median: 1.7%
Percentile 75%: 3.7%
Percentile 90%: 6.7%
Max = 51.2%

Simple Statistics

Variable	N	Mean	Std Dev	Median	Minimum	Maximum
PC_01	1237	0.02753	0.03803	0.01734	0	0.51240
PC_02	1345	0.26287	0.07974	0.25410	0	1.00000
PC_03	162	0.97762	0.03311	0.99425	0.84615	1.00000
PC_04	523	0.05267	0.08432	0.02203	0	0.66129
PC_05	1352	0.49198	0.19284	0.50190	0.00317	1.00000

Spearman Correlation Coefficients Prob > |r| under H0: Rho=0 Number of Observations					
	PC_01	PC_02	PC_03	PC_04	PC_05
PC_01	1.00000 1237	0.06843 0.0163 1231	-0.26960 0.0006 159	0.10724 0.0169 496	-0.03538 0.2137 1237
PC_02	0.06843 0.0163 1231	1.00000 1345	-0.18318 0.0196 162	0.02807 0.5218 523	-0.32009 <.0001 1343
PC_03	-0.26960 0.0006 159	-0.18318 0.0196 162	1.00000 162	-0.03117 0.7030 152	0.07729 0.3283 162
PC_04	0.10724 0.0169 496	0.02807 0.5218 523	-0.03117 0.7030 152	1.00000 523	-0.03560 0.4165 523
PC_05	-0.03538 0.2137 1237	-0.32009 <.0001 1343	0.07729 0.3283 162	-0.03560 0.4165 523	1.00000 1352



The Spearman rank-order correlation is a nonparametric measure of association based on the ranks of the data values by measure PC-01 and hospitals. We used this methodology because of the skewness of the distribution of the measure rates.

2b2.4. What is your interpretation of the results in terms of demonstrating validity? (i.*e., what do the results mean and what are the norms for the test conducted?*)

The correlation of PC-01 with the other PC measures in the PC measure set indicates that the correlations (with the exception of PC-05), although in the expected direction and statistically significant, are relatively weak. Although 90% of the hospital measure rates fall between 0 and 6.7%, there are still a number of hospitals with measure rates significantly greater than 6.7%, indicating that the performance of hospitals on this measure are not uniformly acceptable.

2b3. EXCLUSIONS ANALYSIS .

NA ☐ no exclusions — *skip to section 2b4*

2b3.1. Describe the method of testing exclusions and what it tests (*describe the steps—do not just name a method; what was tested, e.g., whether exclusions affect overall performance scores; what statistical analysis was used*)

There were 1,134,640 admissions selected from the initial cohort. From among the 1,134,640 admissions in 1,237 hospitals, the descriptive statistics are given below.

The following exclusions were analyzed by subpopulation and measure for frequency and variability across providers:

- *ICD-9-CM Principal Diagnosis Code or ICD-9-CM Other Diagnosis Codes* for conditions possibly justifying elective delivery prior to 39 weeks gestation as defined in Appendix A, Table 11.07
- Less than 8 years of age
- Greater than or equal to 65 years of age
- Length of stay > 120 days
- Enrolled in clinical trials
- *Gestational Age* < 37 or >= 39 weeks or UTD

2b3.2. What were the statistical results from testing exclusions? (*include overall number and percentage of individuals excluded, frequency distribution of exclusions across measured entities, and impact on performance measure scores*)

Exclusion Subpopulation 1 - PC-01

ICD-9-CM Principal Diagnosis Code or ICD-9-CM Other Diagnosis Codes for conditions possibly justifying elective delivery prior to 39 weeks gestation as defined in Appendix A, Table 11.07

Exclusion: No observations noted

Less than 8 years of age

Exclusion: Included in the initial population exclusion

Greater than or equal to 65 years of age

Exclusion*:* Included in the initial population exclusion

Length of Stay >120 days

Exclusion: Included in the initial population exclusion

Exclusion: Enrolled in Clinical Trials

Overall Number of Occurrences n = 748

Overall Occurrence Percentage: 0.07%

Minimum: 0%

10th Percentile: 0%

Median: 0%

90th Percentile: 0.062%

Maximum: 28%

Exclusion: Gestational Age < 37 or gestational Age = >39 weeks or UTD

Overall Number of Occurrences n = 851,258

Overall Occurrence Percentage: 84.9%
Minimum 0.29%
10th Percentile: 69.17%
Median: 75.2%
90th Percentile: 79.2%
Maximum: 84.8%

2b3.3. What is your interpretation of the results in terms of demonstrating that exclusions are needed to prevent unfair distortion of performance results? (*i.e., the value outweighs the burden of increased data collection and analysis. Note: **If patient preference is an exclusion**, the measure must be specified so that the effect on the performance score is transparent, e.g., scores with and without exclusion*)

The frequency of exclusions is high for "Gestational Age < 37 or gestational Age = >39 weeks or UTD" Occurrence with an overall percentage equal 75%. The high percentage is justified by the scope of measure PC-01. The difference between the 10th and 90th percentiles of the distribution of exclusion rates is narrow indicating that the occurrence is random and likely would not bias performance results.

It is believed that all of the exclusions should be retained for the following reasons:

Exclusion: *ICD-9-CM Principal Diagnosis Code or ICD-9-CM Other Diagnosis Codes* for conditions possibly justifying elective delivery prior to 39 weeks gestation as defined in Appendix A, Table 11.07
Rationale: Rationale: Table 11.07 contains diagnosis codes for medical conditions that are reasons to perform an early term medical induction and/or cesarean delivery.
Exclusion: Patients who have a *Length of stay* greater than 120 days
Rationale: Included for this measure in order to harmonize with other CMS/Joint Commission aligned measures.

Exclusion: Patients enrolled in a Clinical Trial
Rationale: Only capture patients not enrolled in clinical trials studying pregnant patients or newborns.

Exclusion: Patients with Gestational Age < 37 or >=39 weeks or UTD
Rationale: The denominator population is limited to patients ≥ 37 to < 39 weeks of completed gestation. Patients with UTD for gestational age typically have had no prenatal care.

3. Feasibility

Extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

3a. Byproduct of Care Processes

For clinical measures, the required data elements are routinely generated and used during care delivery (e.g., blood pressure, lab test, diagnosis, medication order).

3a.1. **Data Elements Generated as Byproduct of Care Processes.**

generated by and used by healthcare personnel during the provision of care, e.g., blood pressure, lab value, medical condition, Coded by someone other than person obtaining original information (e.g., DRG, ICD-9 codes on claims), Abstracted from a record by someone other than person obtaining original information (e.g., chart abstraction for quality measure or registry)

If other:

3b. Electronic Sources

The required data elements are available in electronic health records or other electronic sources. If the required data are not in electronic health records or existing electronic sources, a credible, near-term path to electronic collection is specified.

3b.1. **To what extent are the specified data elements available electronically in defined fields?** *(i.e., data elements that are needed to compute the performance measure score are in defined, computer-readable fields)*

Some data elements are in defined fields in electronic sources

3b.2. **If ALL the data elements needed to compute the performance measure score are not from electronic sources, specify a credible, near-term path to electronic capture, OR provide a rationale for using other than electronic sources.**

The Joint Commission recognizes that not all hospitals currently have the capacity to abstract the electronic version of this measure, so continues to offer this chart abstracted version which allows for data capture from unstructured data fields. All data elements needed to compute the PC-01 performance measure score have been retooled for capture from electronic sources. Annual updates are performed to match the eCQM specifications to the current version of the chart-abstracted specifications.

3b.3. **If this is an eMeasure, provide a summary of the feasibility assessment in an attached file or make available at a measure-specific URL.**

Attachment:

3c. Data Collection Strategy

Demonstration that the data collection strategy (e.g., source, timing, frequency, sampling, patient confidentiality, costs associated with fees/licensing of proprietary measures) can be implemented (e.g., already in operational use, or testing demonstrates that it is ready to put into operational use). For eMeasures, a feasibility assessment addresses the data elements and measure logic and demonstrates the eMeasure can be implemented or feasibility concerns can be adequately addressed.

3c.1. **Describe what you have learned/modified as a result of testing and/or operational use of the measure regarding data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, time and cost of data collection, other feasibility/implementation issues.**

IF a PRO-PM, consider implications for both individuals providing PROM data (patients, service recipients, respondents) and those whose performance is being measured.

At the present time, hospitals using this performance measure generally collect measure data via manual review of the paper medical record, the EMR or a combination of both. Collected data are submitted to The Joint Commission on a quarterly basis, by way of contracted performance measurement system vendors, as described previously. Specifications for this measure are freely available to anyone who wishes to use the measure. Feedback from hospitals using this measure indicates that required data elements are generally available in the medical record, and measure specifications are robust and easy to understand. As described above, when feedback from measure users has indicated the need for clarification or revision of measure specifications, this has taken place.

3c.2. **Describe any fees, licensing, or other requirements to use any aspect of the measure as specified** *(e.g., value/code set, risk model, programming code, algorithm)*.

There are no fees or licensing requirements to use the Joint Commission performance measures, all of which are in the public domain.

4. Usability and Use

Extent to which potential audiences (e.g., consumers, purchasers, providers, policy makers) are using or could use performance results for both accountability and performance improvement to achieve the goal of high-quality, efficient healthcare for individuals or populations.

4a. Accountability and Transparency

Performance results are used in at least one accountability application within three years after initial endorsement and are publicly reported within six years after initial endorsement (or the data on performance results are available). If not in use at the time of initial endorsement, then a credible plan for implementation within the specified timeframes is provided.

4.1. Current and Planned Use

NQF-endorsed measures are expected to be used in at least one accountability application within 3 years and publicly reported within 6 years of initial endorsement in addition to performance improvement.

Planned	Current Use (for current use provide URL)
	Public Reporting Quality Check® http://www.qualitycheck.org/consumer/searchQCR.aspx Hospital Compare https://www.medicare.gov/hospitalcompare/search.html Payment Program Hospital Inpatient Quality Reporting Program https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/HospitalQualityInits/HospitalRHQDAPU.html Hospital Value Based Purchasing Program https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/hospital-value-based-purchasing/index.html?redirect=/hospital-value-based-purchasing/ Regulatory and Accreditation Programs Hospital Accreditation Program http://jointcommission.org Quality Improvement with Benchmarking (external benchmarking to multiple organizations) America's Hospitals: Improving Quality and Safety – The Joint Commission's Annual Report http://www.jointcommission.org/annualreport.aspx Quality Improvement (Internal to the specific organization) Perinatal Care Certification http://www.jointcommission.org/certification/perinatal_care_certification.aspx

4a.1. For each CURRENT use, checked above, provide:

- Name of program and sponsor
- Purpose
- Geographic area and number and percentage of accountable entities and patients included

- Name of program and sponsor: Quality Check®; The Joint Commission
- Purpose: A public website that allows consumers to: search for accredited and certified organizations by city and state, by name or by zip code (up to 250 miles); find organizations by type of service provided within a geographic area; download free hospital performance measure results; and, print a list of Joint Commission certified disease-specific care programs and health care staffing firms.
- Geographic area and number and percentage of accountable entities and patients included: Nationwide; 3300 Joint Commission-accredited hospitals (2014)

- Name of program and sponsor: Hospital Compare; Centers for Medicare & Medicaid Services
- Purpose: A public website that provides information to help consumers decide where to obtain healthcare and encourages hospitals to improve the quality of care they provide.
- Geographic area and number and percentage of accountable entities and patients included: Nationwide; 4000+ medicare-certified hospitals (2015)

- Name of program and sponsor: Hospital Inpatient Quality Reporting Program; Centers for Medicare & Medicaid Services
- Purpose: The Hospital Inpatient Quality Reporting (Hospital IQR) program was originally mandated by Section 501(b) of the Medicare Prescription Drug, Improvement, and Modernization Act (MMA) of 2003. This section of the MMA authorized CMS to pay hospitals that successfully report designated quality measures a higher annual update to their payment rates.
- Geographic area and number and percentage of accountable entities and patients included: Nationwide; 3500 hospitals (2015)

- Name of program and sponsor: Hospital Value Based Purchasing Program; Centers for Medicare & Medicaid Services
- Purpose: Hospital Value-Based Purchasing (VBP) is part of the Centers for Medicare & Medicaid Services' (CMS') long-standing effort to link Medicare's payment system to a value-based system to improve healthcare quality, including the quality of care provided in the inpatient hospital setting.
The program attaches value-based purchasing to the payment system that accounts for the largest share of Medicare spending, affecting payment for inpatient stays in over 3,500 hospitals across the country.
Participating hospitals are paid for inpatient acute care services based on the quality of care, not just quantity of the services they provide.
- Geographic area and number and percentage of accountable entities and patients included: Nationwide; 3500 hospitals (2015)

- Name of program and sponsor Hospital Accreditation Program; The Joint Commission
- Purpose: An accreditation program that recognizes hospitals that meet standard requirements to provide safe and effective patient care.
- Geographic area and number and percentage of accountable entities and patients included Nationwide; 3300 Joint Commission-accredited hospitals (2014)

- Name of program and sponsor America's Hospitals: Improving Quality and Safety – The Joint Commission's Annual Report ; The Joint Commission
- Purpose: The annual report summarizes the performance of Joint Commission-accredited hospitals on 46 accountability measures of evidence-based care processes closely linked to positive patient outcomes, and provides benchmarks from Top Performer on Key Quality Measures® hospitals.
- Geographic area and number and percentage of accountable entities and patients included Nationwide; 3300 Joint Commission-accredited hospitals (2014)

- Name of program and sponsor Perinatal Care Certification; The Joint Commission
- Purpose: A certification program that recognizes hospitals that have achieved integrated, coordinated, patient-centered care for clinically uncomplicated pregnancies and births.
- Geographic area and number and percentage of accountable entities and patients included Nationwide; Twelve Joint Commission-accredited hospitals (2016)

4a.2. If not currently publicly reported OR used in at least one other accountability application (e.g., payment program, certification, licensing) what are the reasons? (*e.g., Do policies or actions of the developer/steward or accountable entities restrict access to performance results or impede implementation?*)
Not Applicable

4a.3. If not currently publicly reported OR used in at least one other accountability application, provide a credible plan for implementation within the expected timeframes -- any accountability application within 3 years and publicly reported within 6 years of initial endorsement. (*Credible plan includes the specific program, purpose, intended audience, and timeline for implementing the measure within the specified timeframes. A plan for accountability applications addresses mechanisms for data aggregation and reporting.*)
Not Applicable

4b. Improvement

Progress toward achieving the goal of high-quality, efficient healthcare for individuals or populations is demonstrated. If not in use for performance improvement at the time of initial endorsement, then a credible rationale describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.

4b.1. **Progress on Improvement. (Not required for initial endorsement unless available.)**
Performance results on this measure (current and over time) should be provided in 1b.2 and 1b.4. Discuss:

- **Progress (trends in performance results, number and percentage of people receiving high-quality healthcare)**
- **Geographic area and number and percentage of accountable entities and patients included**

The rate of early term elective deliveries has decreased from 18.8% in 2010 with 164 hospitals reporting to 3.3% in 2014 with 1388 hospitals reporting based on Joint Commission ORYX performance measurement data. Beginning with January 1, 2016 discharges, an additional 821 accredited hospitals will begin reporting the data. The new reporting requirement will capture approximately 80% of the accredited hospitals with maternity services in the US. Additionally, the adoption of PC-01 into the CMS Hospital Inpatient Quality Reporting (HIQR) Program has significantly increased the number of hospitals reporting their early term elective delivery data with 3500 hospitals participating in the reporting program.

4b.2. **If no improvement was demonstrated, what are the reasons? If not in use for performance improvement at the time of initial endorsement, provide a credible rationale that describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.**
Not Applicable

4c. Unintended Consequences

The benefits of the performance measure in facilitating progress toward achieving high-quality, efficient healthcare for individuals or populations outweigh evidence of unintended negative consequences to individuals or populations (if such evidence exists).

4c.1. **Were any unintended negative consequences to individuals or populations identified during testing; OR has evidence of unintended negative consequences to individuals or populations been reported since implementation? If so, identify the negative unintended consequences and describe how benefits outweigh them or actions taken to mitigate them.**
Unintended Consequence:
Cases with prior uterine surgery were inappropriately failing the measure.
Mitigating Action:
The measure rate calculation algorithm was revised to include a check prior to a cesarean birth via a new data element (Prior Uterine Surgery) created to enable cases with prior uterine surgery to remain in the denominator population and pass the measure.

Unintended Consequence:
Patients who did not receive prenatal care were inappropriately included in the measure denominator, as the gestational age data element was abstracted as unable to be determined (UTD).
Mitigating Action:
In order to avoid penalizing hospitals, cases with UTD were removed from the measure population.

Unintended Consequence:
Some hospitals have reported higher rates due to small denominator populations as a result of sampling.
Mitigating Action:
Vital Records reports, delivery logs and clinical information systems were added as acceptable data sources to help hospitals identify all cases with 37 and 38 weeks gestation, so that 100% of these cases could be reviewed to increase the denominator population size.

5. Comparison to Related or Competing Measures

If a measure meets the above criteria and there are endorsed or new related measures (either the same measure focus or the same target population) or competing measures (both the same measure focus and the same target population), the measures are compared to address harmonization and/or selection of the best measure.

5. Relation to Other NQF-endorsed Measures
Are there related measures (conceptually, either same measure focus or target population) or competing measures (conceptually both the same measure focus and same target population)? If yes, list the NQF # and title of all related and/or competing measures.
No

5.1a. **List of related or competing measures (selected from NQF-endorsed measures)**

5.1b. **If related or competing measures are not NQF endorsed please indicate measure title and steward.**

5a. Harmonization

The measure specifications are harmonized with related measures;
OR
The differences in specifications are justified

5a.1. **If this measure conceptually addresses EITHER the same measure focus OR the same target population as NQF-endorsed measure(s):**
Are the measure specifications completely harmonized?

5a.2. **If the measure specifications are not completely harmonized, identify the differences, rationale, and impact on interpretability and data collection burden.**
Not Applicable

5b. Competing Measures

The measure is superior to competing measures (e.g., is a more valid or efficient way to measure);
OR
Multiple measures are justified.

5b.1. **If this measure conceptually addresses both the same measure focus and the same target population as NQF-endorsed measure(s):**
Describe why this measure is superior to competing measures (e.g., a more valid or efficient way to measure quality); OR provide a rationale for the additive value of endorsing an additional measure. (Provide analyses when possible.)
Not Applicable

Appendix

A.1 Supplemental materials may be provided in an appendix. All supplemental materials (such as data collection instrument or methodology reports) should be organized in one file with a table of contents or bookmarks. If material pertains to a specific submission form number, that should be indicated. Requested information should be provided in the submission form and required attachments. There is no guarantee that supplemental materials will be reviewed.

Available at measure-specific web page URL identified in S.1 **Attachment:**

Contact Information

Co.1 **Measure Steward (Intellectual Property Owner):** The Joint Commission
Co.2 **Point of Contact:** Ann, Watt, awatt@jointcommission.org, 630-792-5944-
Co.3 **Measure Developer if different from Measure Steward:** The Joint Commission
Co.4 **Point of Contact:** Ann, Watt, awatt@jointcommission.org, 630-792-5944-

Additional Information

Ad.1 **Workgroup/Expert Panel involved in measure development**

Provide a list of sponsoring organizations and workgroup/panel members' names and organizations. Describe the members' role in measure development.

Michael Ross, MD, MPH (Chair)
Harbor-UCLA Medical Center
Torrance, CA

Wanda Barfield, MD, MPH
Centers for Disease Control and Prevention
Atlanta, GA

Kenneth E. Brown, MD, MBA, FACOG, FACHE
Woman´s Hospital
Lafayette, LA

Martin McCaffrey, MD
UNC North Carolina Children's Hospital
Chapel Hill, NC

Cathy Collins-Fulea, MSN, CNM
Henry Ford Hospital
Detroit, MI

Janet H. Muri, MBA
National Perinatal Information Center/
Quality Analytic Services
Providence, RI

Kathleen Simpson, PhD, RNC, FAAN
St. John's Mercy Medical Center
St. Louis, MO

Michael Socol, MD
Northwestern University Medical School
Chicago, IL

Rebecca Zimmermann, MPP
America's Health Insurance Plans
Washington, DC

The technical advisory panel (TAP) members determined priority areas that could be evaluated to improve care related to perinatal care during the development timeframe. After implementation, minor revisions, acknowledged by TAP representatives, were made to improve clarity. Hospital feedback will be reviewed during the reliability testing phase of the project to assist the TAP in making the final measure recommendations.

Measure Developer/Steward Updates and Ongoing Maintenance

Ad.2 **Year the measure was first released:** 2010
Ad.3 **Month and Year of most recent revision:** 10, 2015
Ad.4 **What is your frequency for review/update of this measure?** Biannual
Ad.5 **When is the next scheduled review/update for this measure?** 02, 2016

Ad.7 **Disclaimers:**
Ad.8 **Additional Information/Comments:**

MEASURE WORKSHEET

This document summarizes the evaluation of the measure as it progresses through NQF's Consensus Development Process (CDP). The information submitted by measure developers/stewards is included after the Brief Measure Information, Preliminary Analysis, and Pre-meeting Public and Member Comments sections.

To navigate the links in the worksheet: Ctrl + click link to go to the link; ALT + LEFT ARROW to return

Brief Measure Information

NQF #: 0470
Measure Title: Incidence of Episiotomy
Measure Steward: Christiana Care Health System
Brief Description of Measure: Percentage of vaginal deliveries (excluding those coded with shoulder dystocia) during which an episiotomy is performed.
Developer Rationale: Episiotomy has been clearly linked with worse perineal tears and in turn its attendant complications. These are noted to include perineal pain, blood loss, and potential for wound break down/abscess formation and necrotizing fascitis. Predicated on these concerns, ACOG has called for "restricted use of episiotomy".

Numerator Statement: Number of episiotomy procedures (ICD-9 code 72.1, 72.21, 72.31, 72.71, 73.6; **ICD-10 PCS:0W8NXZZ** performed on women undergoing a vaginal delivery (excluding those with shoulder dystocia **ICD-10; O66.0**) during the analytic period- monthly, quarterly, yearly etc.
Denominator Statement: All vaginal deliveries during the analytic period- monthly, quarterly, yearly etc. excluding those coded with a shoulder dystocia **ICD-10: O66.0**).
Denominator Exclusions: Women who have a coded complication of shoulder dystocia. In the case of shoulder dystocia, an episiotomy is performed to free the shoulder and prevent/mitigate birth injury to the infant.

Measure Type: Process
Data Source: Administrative claims, Paper Medical Records
Level of Analysis: Facility

IF Endorsement Maintenance – Original Endorsement Date: Oct 24, 2008 **Most Recent Endorsement Date:** Apr 02, 2012

Maintenance of Endorsement -- Preliminary Analysis

To maintain NQF endorsement endorsed measures are evaluated periodically to ensure that the measures still meets the NQF endorsement criteria ("maintenance"). The emphasis for maintaining endorsement is focused on how effective the measure is for promoting improvements in quality. Endorsed measures should have some experience from the field to inform the evaluation. The emphasis for maintaining endorsement is noted for each criterion.

Criteria 1: Importance to Measure and Report

1a. Evidence
Maintenance measures – less emphasis on evidence unless there is new information or change in evidence since the prior evaluation.

1a. Evidence. The evidence requirements for a *process or intermediate outcome* measure is that it is based on a systematic review (SR) and grading of the body of empirical evidence where the specific focus of the evidence matches what is being measured.

The developer provides the following evidence for this measure:

- **Systematic Review of the evidence specific to this measure?** ☒ **Yes** ☐ **No**

- **Quality, Quantity and Consistency of evidence provided?** ☒ **Yes** ☐ **No**
- **Evidence graded?** ☒ **Yes** ☐ **No**

Summary of prior review in 2016

- ACOG Practice Bulletin no. 71 (April 2006) cites the "restricted use of episiotomy is preferable to routine use" as level A evidence. The developer indicates that "The evidence is direct in that restricted use of episiotomy has been firmly linked to lower rates of perineal injury. Thus decreasing the routine use of episiotomy one can directly influence the rate of perineal injury."

Changes to evidence from last review

☒ **The developer attests that there have been no changes in the evidence since the measure was last evaluated.**

☐ **The developer provided updated evidence for this measure:**

Exception to evidence NA

Guidance from the Evidence Algorithm

Process measure (Box 1) → Systematic review (Box 3) → QQC present (Box 4) →SR concludes high quality evidence

Questions for the Committee:

- The developer attests the underlying evidence for the measure has not changed since the last NQF endorsement review. Does the Committee agree the evidence basis for the measure has not changed and there is no need for repeat discussion and voting on Evidence?

Preliminary rating for evidence: ☒ **Pass** ☐ **No Pass**

1b. Gap in Care/Opportunity for Improvement and 1b. disparities
Maintenance measures – increased emphasis on gap and variation

1b. Performance Gap. The performance gap requirements include demonstrating quality problems and opportunity for improvement.

The developer provided the following data:

National Quality Forum (NQF)
Incidence of Episiotomy for Vaginal Deliveries without Shoulder Dystocia
2010 - 2014

n = 63 hospitals in NPIC/QAS Trend database	2010	2011	2012	2013	2014	Pct Change 2010 - 2014
Total vaginal deliveries without shoulder dystocia	168,949	169,729	169,042	167,941	169,181	0.1%
Vaginal deliveries without shoulder dystocia with episiotomy	17,808	16,153	14,803	13,231	11,862	-33.4%
Percent (Trend Hospitals weighted average) *	10.5%	9.5%	8.8%	7.9%	7.0%	-33.5%
Percent (Trend Hospitals unweighted average) *	11.3%	10.3%	9.5%	8.4%	7.3%	-35.9%
Range						
Low	0.6%	0.7%	0.6%	0.8%	0.8%	
High	35.1%	32.8%	29.3%	25.2%	22.1%	

* Trend rate (Weighted and Unweighted): Significant Downward Trend, p = 0.000

Disparities

Specific disparities information was not provided other than the variation in rates from 0.8% to 22.1%.

Questions for the Committee:

- *IS there opportunity for improvement? Is there a gap in care that warrants a national performance measure?*
- *No disparities information is provided, are you aware of evidence that disparities exist in this area of healthcare?*

Preliminary rating for opportunity for improvement: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 1: Importance to Measure and Report (including 1a, 1b, 1c)

1a. Evidence to Support Measure Focus

Comments:

While evidence has not recently changed for episiotomy, the evidence is strong that episiotomy leads to harm by worsening perineal tears. There are appropriate exclusions for should dystocia.

Episiotomies are directly linked to an increased risk of perineal injury. I would argue that an episiotomy is an injury to the perineum.

The evidence for this intermediate outcome measure is strong and the avoidance of "routine" episiotomy during vaginal birth has been recommended by ACOG since 2006. The most recent Cochrane review is 2009.

1b. Performance Gap

Comments:

**Performance gap is demonstrated by large variation in episiotomy rates as well as the fact that overall rates are higher than what might be considered ideal for a process measure. **

Data was provided and there is still an opportunity for improvement. This opportunity still warrants a national measure. Data on the measure by population subgroup was not provided, and that would be nice to see if it's available.

Yes, from 63 hospitals in NPIC/QAS (National Perinatal Information Center/Quality Analytic Services) shows a 33.4% decline from 2010-2014 in the incidence of total vaginal deliveries without shoulder dystocia with episiotomy. The remarkable issue is the wide variation among hospitals from an incidence of episiotomy of 0.8% to 22.1% in 2014. These data alone point to the continued need for endorsement of this measure. No racial/ethnic or SES disparities information was provided. The disparity that I am aware of is hospital to hospital rather than population specific disparities. Although the NQF reviewer noted that the preliminary rating for opportunity for improvement was "Moderate", I would rate it as "High" due to the reported hospital to hospital variation.

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability

2a1. Reliability Specifications

Maintenance measures – no change in emphasis – specifications should be evaluated the same as with new measures

2a1. Specifications requires the measure, as specified, to produce consistent (reliable) and credible (valid) results about the quality of care when implemented.

Data source(s): Administrative claims, paper medical records

Specifications:

- Specifications are clearly defined. The numerator is the number of episiotomies performed on women undergoing a vaginal delivery (excluding those with shoulder dystocia) and the denominator is all vaginal deliveries within the analytic period.
- The level of analysis is facility level.

- The measure is updated to include ICD-10 codes.
- The measure is not risk adjusted.
- The developer reports that there have been no changes to the specification except for updating to ICD-10 codes as of 2015.

Questions for the Committee:

- *Are all the data elements clearly defined? Are all appropriate codes included?*
- *Is the logic or calculation algorithm clear?*
- *Is it likely this measure can be consistently implemented?*

2a2. Reliability Testing Testing attachment
Maintenance measures – less emphasis if no new testing data provided

2a2. Reliability testing demonstrates if the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise enough to distinguish differences in performance across providers.

For maintenance measures, summarize the reliability testing from the prior review:

- The developers provided testing of data element validity (comparing the coded data to the medical record "gold standard")

SUMMARY OF TESTING

Reliability testing level ☐ **Measure score** ☒ **Data element** ☐ **Both**

Reliability testing performed with the data source and level of analysis indicated for this measure ☒ **Yes** ☐ **No**

Method(s) of reliability testing Data element validity

Results of reliability testing see validity below

Guidance from the Reliability Algorithm

Precise specifications (Box 1) →No empirical reliability testing of the score (Box 2) → empirical validity testing using patient-level data → Use data element validity testing for patient-level data elements (see below)

The rating for validity will count for reliability also.

2b. Validity
Maintenance measures – less emphasis if no new testing data provided

2b1. Validity: Specifications

2b1. Validity Specifications. This section should determine if the measure specifications are consistent with the evidence.

Specifications consistent with evidence in 1a. ☒ **Yes** ☐ **Somewhat** ☐ **No**

Question for the Committee:

- *Are the specifications consistent with the evidence?*

2b2. Validity testing

2b2. Validity Testing should demonstrate the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality.

SUMMARY OF TESTING

Validity testing level ☐ **Measure score** ☒ **Data element testing against a gold standard** ☐ **Both**

Method of validity testing of the measure score:

- ☒ **Face validity only**
- ☐ **Empirical validity testing of the measure score**

Validity testing method: Two methods of validity testing are presented:

1. The hospitals being measured compared the coded data with the medical record ("gold standard").
2. Face validity was assessed by the hospitals being measured. In addition to auditing their sample of cases, hospitals were asked to rate the reliability, validity, feasibility and usability of this measure

Validity testing results:

1. Hospital provided data on comparison with coded data and the medical record:

11 Hospitals	# deliveries	# episiotomy cases	Coded correctly	Incorrect coding
2008-2009 (1 year)	66,306	9626	7 of 9 (63%)	1 hospital: 1 case miscoded 1 hospital: no response
2009-2010 (6 months)	31,496	4259	4 of 11 (36.6%)	7 hospitals: 1-4 cases miscoded evenly divided between those with and without episiotomy

2. Face validity testing was performed. In 2010, 9 of 11 hospitals (81.8%) indicated they felt episiotomy rate is a valid measure of the quality of care at a hospital; the other 2 felt the measure was valid but should be looked at at the provider level since providers will determine whether to perform an episiotomy or not.

2b3-2b7. Threats to Validity

2b3. Exclusions:

- Expert opinion recommends episiotomy in the case of shoulder dystocia to free the shoulder and prevent injury to the infant. The estimated incidence is 1%. Cases of shoulder dystocia are excluded.

Questions for the Committee:

- *Are the exclusions consistent with the evidence?*
- *Are any patients or patient groups inappropriately excluded from the measure?*
- *Are the exclusions/exceptions of sufficient frequency and variation across providers to be needed (and outweigh the data collection burden)?*

2b4. Risk adjustment: **Risk-adjustment method** ☒ **None** ☐ **Statistical model** ☐ **Stratification**

2b5. Meaningful difference *(can statistically significant and clinically/practically meaningful differences in performance measure scores can be identified)*:

- In the 12 facilities, rates ranged from 33.1% to 3.4%, indicating a wide variation in performance. The unweighted average is 14.9% and the weighted average is 13.%.

Question for the Committee:

- *Does this measure identify meaningful differences about quality?*

2b6. Comparability of data sources/methods: Not applicable.

2b7. Missing Data

- No information was provided.

Questions for the Committee:

- *The developer attests that no additional testing has been performed since the last NQF endorsement review. Does the Committee agree the testing data is sufficient and there is no need for repeat discussion and voting on reliability and validity?*

Preliminary rating for validity: ☒ Pass ☐ No Pass ☐ Insufficient

Committee pre-evaluation comments

Criteria 2: Scientific Acceptability of Measure Properties (including all 2a, 2b, and 2d)

2a1. & 2b1. Specifications

Comments:

No concerns.

Clearly defined elements. All appropriate codes are included. The algorithm is clear. No concerns regarding implementation.

The specifications are clear and available in either EHR or paper medical records. ICD-10 codes updated. Numerator = # of episiotomies performed on women experiencing vaginal birth EXCLUDING those with dx shoulder dystocia.Denominator - All vaginal deliveries within the periodFacility level reporting (although improvement could be facilitated by individual provider level reporting)I see no issues with the logic/calculation and am confident the measure can be consistently implemented.

Measure is meaningful in that it is strongly related to an important health outcome and under the control of the health care system/provider.

Consistent with evidence.

There are no elements of the specifications that are inconsistent with the evidence. Shoulder dystocia remains a recognized clinical indication for episiotomy, so is appropriately excluded.

2a2. Reliability Testing

Comments:

Testing was done and there seems to be high reliability.

Yes, testing was provided of the data element validity.

Reliability is not an issue of measure specification, but the accuracy of the medical record (EHR or paper) and the accuracy of data abstraction (inter-rater reliability). The first issue is always present in any data from the medical record; the second was described under validity that 7 of 9 hospitals coded all episiotomies correctly. Periodic errors were evident.

2b2. Validity Testing

Comments:

Both face and comparison of coding with medical record. Both types of validity did well.

Adequate scope. I agree that the episiotomy rate should be reviewed at the facility and provider level.

Evidence from face validity and accuracy of hospital coding.

2b3. Exclusions Analysis

2b4. Risk Adjustment/Stratification for Outcome or Resource Use Measures

2b5. Identification of Statistically Significant & Meaningful Differences In Performance

2b6. Comparability of Performance Scores When More Than One Set of Specifications

2b7. Missing Data Analysis and Minimizing Bias

Comments:

**Data can come from multiple data sources and still be valid. No risk adjustment is needed. Data suggest wide range of rates thus improvements could be meaningful.

while large amounts of missing data could threaten this measure. This is unlikely given the data sources. There is some evidence that codes can vary by insurance status and demographics, but this is likely undercoding and not overcoding. Thus high rates are still problematic.**

Have enough of a sample size this does not pose a threat.
**Episiotomies for shoulder dystocia excluded and this is consistent with the evidence. No risk adjustment recommended or needed. The data in 1.b re: the performance gap indicates there remains meaningful differences in U.S. maternity care that merit attention. Missing data are probably less likely for this measure than others at the time of birth since it is a response category on the EHR or paper record. The primary threat to validity is when the birth record does not match the provider's delivery note since these are usually completed by different individuals (nurse and provider). **

Criterion 3. Feasibility
Maintenance measures – no change in emphasis – implementation issues may be more prominent

3. Feasibility is the extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

- All data elements are in defined electronic fields. The measure uses administrative claims which is generally considered feasible.

Questions for the Committee:

- *Are the required data elements routinely generated and used during care delivery?*

Preliminary rating for feasibility: ☒ **High** ☐ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 3: Feasibility

3a. Byproduct of Care Processes
3b. Electronic Sources
3c. Data Collection Strategy

Comments:
Very feasible to use. Data sources are cheap and widely available. Hand collected data could also be used if someone wanted to adapt the measure to a data set with high quality than administrative data.
Elements should be available. No concerns.
This is a very feasible measure since it is a standard element of the delivery record. For the EHR, data abstraction should require minimal effort. Once the algorithm is established in the retrieval strategy to eliminate vaginal births with shoulder dystocia, both numerator and denominator are easily obtained and the statistic calculated. No concerns re: the operationalization of the data collection strategy.

Criterion 4: Usability and Use
Maintenance measures – increased emphasis – much greater focus on measure use and usefulness, including both impact /improvement and unintended consequences

4. Usability and Use evaluate the extent to which audiences (e.g., consumers, purchasers, providers, policymakers) use or could use performance results for both accountability and performance improvement activities.

Current uses of the measure
Publicly reported? ☐ **Yes** ☐ **No**

Staff note: The Leapfrog Group publishes hospital results of several maternity care measures including incidence of episiotomy. The data is voluntarily submitted by hospitals and reported on Leapfrog's website at http://www.leapfroggroup.org/cp?frmbmd=cp_listings&find_by=state&city=&state=DC . Leapfrog has set a target rate of 5%.

Current use in an accountability program? ☒ **Yes** ☐ **No**

OR

Planned use in an accountability program? ☒ **Yes** ☐ **No**

Accountability program details

- National Perinatal Information Center, Inc. (NPIC) Metric reported to member and military hospitals for quality review and bench marking. Total 400,000+ deliveries across all 130 hospitals and the majority of states.

Improvement results

- The NPIC database shows a drop in episiotomy rates over time. The unweighted average (preferred method for hospital to hospital comparisons) went from 11.3% in 2010 to 7.3% in 2014. The range in 2014 was .8% to 22.1 % indicating continued opportunities for improvement.

Unexpected findings (positive or negative) during implementation

- No information provided.

Potential harms

- No information provided.

Feedback:

- No specific feedback on this measure submitted to NQF.

***Questions for the Committee**:*

- *How can the performance results be used to further the goal of high-quality, efficient healthcare?*
- *Do the benefits of the measure outweigh any potential unintended consequences?*

Preliminary rating for usability and use: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 4: Usability and Use

4a. Accountability and Transparency

4b. Improvement

4c. Unintended Consequences

Comments:

Measure is currently being used and data show that rates are improving (dropping).

Reported voluntarily on LeapFrog. Consider it being a TJC measure in the future and require mandatory reporting. No unintended consequences.

The Leapfrog Group publishes hospital level episiotomy incidences. Data is voluntarily submitted by hospitals. Leapfrog target is 5%. NPIC metric for member and military hospitals for quality review. Is not included in the Joint Commission's Perinatal Core Measures.

Criterion 5: Related and Competing Measures

Related or competing measures

- There are no related or competing measures.

Harmonization

- N/A

Pre-meeting public and member comments

-

1. IMPACT, OPPORTUITY, EVIDENCE - IMPORTANCE TO MEASURE AND REPORT

Importance to Measure and Report is a threshold criterion that must be met in order to recommend a measure for endorsement. All three subcriteria must be met to pass this criterion. See guidance on evidence.
Measures must be judged to be important to measure and report in order to be evaluated against the remaining criteria. **(evaluation criteria)**

1c.1 **Structure-Process-Outcome Relationship** *(Briefly state the measure focus, e.g., health outcome, intermediate clinical outcome, process, structure; then identify the appropriate links, e.g., structure-process-health outcome; process- health outcome; intermediate clinical outcome-health outcome)*:
This is a process measure.

1c.2-3 **Type of Evidence** *(Check all that apply)*:
Clinical Practice Guideline

1c.4 **Directness of Evidence to the Specified Measure** *(State the central topic, population, and outcomes addressed in the body of evidence and identify any differences from the measure focus and measure target population)*:
The evidence is direct in that restricted use of episiotomy has been firmly linked to lower rates of perineal injury. Thus decreasing the routine use of episiotomy one can directly influence the rate of perineal injury. This would apply to all women delivering vaginally and thus the there are no differences between the measure focus and target population.

1c.5 **Quantity of Studies in the Body of Evidence** *(Total number of studies, not articles)*: A pubmed search, reveals 2160 articles on episiotomy of which 195 are reviews.

1c.6 **Quality of Body of Evidence** (*Summarize the certainty or confidence in the estimates of benefits and harms to patients across studies in the body of evidence resulting from study factors. Please address: a) study design/flaws; b) directness/indirectness of the evidence to this measure (e.g., interventions, comparisons, outcomes assessed, population included in the evidence); and c) imprecision/wide confidence intervals due to few patients or events):* In its review of the subject, ACOG cites the "restricted use of episiotomy is preferable to routine use" as level A evidence. The research is too broad to address failings and limitations of individual studies.

1c.7 **Consistency of Results across Studies** *(Summarize the consistency of the magnitude and direction of the effect):* See above

1c.8 **Net Benefit** *(Provide estimates of effect for benefit/outcome; identify harms addressed and estimates of effect; and net benefit - benefit over harms)*:
The lowest achievable rate of episiotomy remains unclear. Nonetheless as previously stated in our internal review 6-7%% of women continue to undergo this procedure in 2014. The exact percentage of women who would directly benefit beyond avoidance of this procedure remains unclear.

1c.9 **Grading of Strength/Quality of the Body of Evidence**. Has the body of evidence been graded? Yes

1c.10 **If body of evidence graded, identify the entity that graded the evidence including balance of representation and any disclosures regarding bias:** This was graded by ACOG.

1c.11 System Used for Grading the Body of Evidence: GRADE

1c.12 If other, identify and describe the grading scale with definitions:

1c.13 Grade Assigned to the Body of Evidence: A

1c.14 Summary of Controversy/Contradictory Evidence: None

1c.15 Citations for Evidence other than Guidelines*(Guidelines addressed below)*:

Moore, E. & Moorhead, C. Promoting normality in the management of the perineum during the second stage of labor. British Journal of Midwifery. September 2013, 21 (9):616-620.

Agency for Healthcare Research and Quality. Use of Episiotomy and Forceps During Childbirth Down, C-Section Rates Up. AHRQ News and Numbers. April 28, 2011.Retrieved from http://archive.ahrq.gov/news/newsroom/news-and-numbers/042811.html

Center for Disease Control. National Vital Statistics Reports, June 28, 2013, 62 (1): 11.

Center for Disease Control. Number, rate and standard error of all-listed surgical and nonsurgical procedures for discharges from short stay hospitals by selected categories; United States, 2009. Retrieved from www.cdc.gov/nchs/data/nhds/4procedure/2009pro4_memberrate.pdf

The American College of Obstetrics and Gynecologists. Limitations of Perineal lacerations as an Obstetrical Quality measure. Committee Opinion. November, 2015, 647.

Stedenfeldt, M. et al. Anal incontinence, urinary incontinence and sexual problems in primiparous women- a comparison between women with episiotomy only and women with episiotomy and obstetric anal sphincter injury. BMC Women's Health, 2014, 14: 157. Doi:10 1186/s12905-041-015-y

Op.cit, Moore & Moorhead

Op.cit. American College of Obstetricians and Gynecologists. Committee Opinion

Martin JA, Hamilton BE, Ventura SJ, Menacker F, Park MM. Births: final data for 2000. Natl Vital Stat Rep 2002;50(5):1–101. (Level II-3)

DeLee JB. The prophylactic forceps operation. Am J Obstet Gynecol 1920;1:34–44. (Level III)

Pomeroy RH. Shall we cut and reconstruct the perineum for every primipara? Am J Obstet Dis Women Child 1918;78:211–20. (Level III)

Thacker SB, Banta HD. Benefits and risks of episiotomy: an interpretive review of the English language literature, 1860-1980. Obstet Gynecol Surv 1983;38:322–38. (Level III)

DeFrances CJ, Hall MJ, Podgornik MN. 2003 National Hospital Discharge Survey. Advance data; No. 359. Hyattsville (MD): National Center for Health Statistics; 2005. Available at: http://www.cdc.gov/nchs/data/ad/ad359.pdf. Retrieved December 29, 2005. (Level II-3)

Martin JA, Hamilton BE, Sutton PD, Ventura SJ, Menacker F, Munson ML. Births: final data for 2003. Natl Vital Stat Rep 2005;54(2):1–116. (Level II-3)

Coats PM, Chan KK, Wilkins M, Beard RJ. A comparison between midline and mediolateral episiotomies. Br J Obstet Gynaecol 1980;87:408–12. (Level II-1)

Bodner-Adler B, Bodner K, Kaider A, Wagenbichler P, Leodolter S, Husslein P, et al. Risk factors for third-degree perineal tears in vaginal delivery, with an analysis of episiotomy types. J Reprod Med 2001;46:752–6. (Level II-3)

Riskin-Mashiah S, O´Brian Smith E, Wilkins IA. Risk factors for severe perineal tear: can we do better? Am J Perinatol 2002;19:225–

34. (Level II-2)

Helwig JT, Thorp JM Jr, Bowes WA Jr. Does midline episiotomy increase the risk of third- and fourth-degree lacerations in operative vaginal deliveries? Obstet Gynecol 1993;82:276–9. (Level II-2)

Shiono P, Klebanoff MA, Carey JC. Midline episiotomies: more harm than good? Obstet Gynecol 1990;75:765–70. (Level II-2)

Oboro VO, Tabowei TO, Loto OM, Bosah JO. A multicentre evaluation of the two-layered repair of postpartum perineal trauma. J Obstet Gynaecol 2003;23:5–8. (Level I)

Grant A, Gordon B, Mackrodt C, Fern E, Truesdale A, Ayers S. The Ipswich childbirth study: one year followup of alternative methods used in perineal repair. BJOG 2001;108:34–40. (Level II-2)

Gordon B, Mackrodt C, Fern E, Truesdale A, Ayers S, Grant A. The Ipswich Childbirth Study: I. A randomised evaluation of two stage postpartum perineal repair leaving the skin unsutured. Br J Obstet Gynaecol 1998;105: 435–40. (Level I)

Kettle C, Hills RK, Jones P, Darby L, Gray R, Johanson R. Continuous versus interrupted perineal repair with standard or rapidly absorbed sutures after spontaneous vaginal birth: a randomised controlled trial. Lancet 2002;359:2217–23. (Level I)

Mahomed K, Grant A, Ashurst H, James D. The Southmead perineal suture study. A randomized comparison of suture materials and suturing techniques for repair of perineal trauma. Br J Obstet Gynaecol 1989;96:1272–80. (Level I)

Mackrodt C, Gordon B, Fern E, Ayers S, Truesdale A, Grant A. The Ipswich Childbirth Study: 2. A randomised comparison of polyglactin 910 with chromic catgut for postpartum perineal repair. Br J Obstet Gynaecol 1998;105:441–5. (Level I)

Grant A. The choice of suture materials and techniques for repair of perineal trauma: an overview of the evidence from controlled trials. Br J Obstet Gynaecol 1989;96:1281–9. (Level III)

Ketcham KR, Pastorek JG 2nd, Letellier RL. Episiotomy repair: chromic versus polyglycolic acid suture. South Med J 1994;87:514–7. (Level III)

Hankins GD, Hauth JC, Gilstrap LC 3rd, Hammond TL, Yeomans ER, Snyder RR. Early repair of episiotomy dehiscence. Obstet Gynecol 1990;75:48–51. (Level III)

Barranger E, Haddad B, Paniel BJ. Fistula in ano as a rare complication of mediolateral episiotomy: report of three cases. Am J Obstet Gynecol 2000;182:733–4. (Level III)

Myles TD, Santolaya J. Maternal and neonatal outcomes in patients with prolonged second stage of labor. Obstet Gynecol 2003;102:52–8. (Level II-3)

Bodner-Adler B, Bodner K, Kimberger O, Wagenbichler P, Mayerhofer K. Management of the perineum during forceps delivery. Association of episiotomy with the frequency and severity of perineal trauma in women undergoing forceps delivery. J Reprod Med 2003;48:239–42. (Level II-3)

Hartmann K, Viswanathan M, Palmieri R, Gartlehner G, Thorp J, Lohr KN. Outcomes of routine episiotomy: a systematic review. JAMA 2005;293:2141–8. (Level III)

Eason E, Labrecque M, Wells G, Feldman P. Preventing perineal trauma during childbirth: a systematic review. Obstet Gynecol 2000;95:464–71. (Meta-Analysis)

Fenner DE, Genberg B, Brahma P, Marek L, DeLancey JO. Fecal and urinary incontinence after vaginal delivery with anal sphincter disruption in an obstetrics unit in the United States. Am J Obstet Gynecol 2003;189:1543–50. (Level II-3)

Robinson JN, Norwitz ER, Cohen AP, McElrath TF, Lieberman ES. Epidural analgesia and third- and fourth-degree lacerations in nulliparas. Obstet Gynecol 1999;94:259–62. (Level II-3)

Sartore A, De Seta F, Maso G, Pregazzi R, Grimaldi E, Guaschino S. The effects of mediolateral episiotomy on pelvic floor function after vaginal delivery. Obstet Gynecol 2004;103:669–73. (Level II-2)

MacArthur C, Bick DE, Keighley MR. Faecal incontinence after childbirth. Br J Obstet Gynaecol 1997;104:46–50.

Walsh CJ, Mooney EF, Upton GJ, Motson RW. Incidence of third-degree perineal tears in labour and outcome after primary repair. Br J Surg 1996;83:218–21. (Level II-2)

Fleming N, Newton ER, Roberts J. Changes in postpartum perineal muscle function in women with and without episiotomies. J Midwifery Womens Health 2003;48:53–9. (Level II-2)

Thranov I, Kringelbach AM, Melchior E, Olsen O, Damsgaard MT. Postpartum symptoms. Episiotomy or tear at vaginal delivery. Acta Obstet Gynecol Scand 1990;69:11–5. (Level II-3)

Macarthur AJ, Macarthur C. Incidence, severity, and determinants of perineal pain after vaginal delivery: a prospective cohort study. Am J Obstet Gynecol 2004;191:1199–204. (Level II-2)

Signorello LB, Harlow BL, Chekos AK, Repke JT. Postpartum sexual functioning and its relationship to perineal trauma: a retrospective cohort study of primiparous women. Am J Obstet Gynecol 2001;184:881–7; discussion 888–90. (Level II-2)

Abraham S, Child A, Ferry J, Vizzard J, Mira M. Recovery after childbirth: a preliminary prospective study. Med J Aust 1990;152:9–12. (Level II-2)

Isager-Sally L, Legarth J, Jacobsen B, Bostofte E. Episiotomy repair—immediate and long-term sequelae. A prospective randomized study of three different methods of repair. Br J Obstet Gynaecol 1986;93:420–5. (Level I)

Upton A, Roberts CL, Ryan M, Faulkner M, Reynolds M, Raynes-Greenow C. A randomised trial, conducted by midwives, of perineal repairs comparing a polyglycolic suture material and chromic catgut. Midwifery 2002;18:223–9. (Level I)

Bowen ML, Selinger M. Episiotomy closure comparing enbucrilate tissue adhesive with conventional sutures. Int J Gynaecol Obstet 2002;78:201–5. (Level II-1)

Nocon JJ, McKenzie DK, Thomas LJ, Hansell RS. Shoulder dystocia: an analysis of risks and obstetric maneuvers. Am J Obstet Gynecol 1993;168:1732–7; discussion 1737–9. (Level II-3)

Rockner G, Wahlberg V, Olund A. Episiotomy and perineal trauma during childbirth. J Adv Nurs 1989;14:264–8. (Level II-2)

Poen AC, Felt-Bersma RJ, Dekker GA, Deville W, Cuesta MA, Meuwissen SG. Third degree obstetric perineal tears: risk factors and the preventive role of mediolateral episiotomy. Br J Obstet Gynaecol 1997;104:563–6. (Level II-2)

Signorello LB, Harlow BL, Chekos AK, Repke JT. Midline episiotomy and anal incontinence: a retrospective cohort study. BMJ 2000;320:86–90. (Level II-2)

De Leeuw JW, Vierhout ME, Struijk PC, Hop WC, Wallenburg HC. Anal sphincter damage after vaginal delivery: functional outcome and risk factors for fecal incontinence. Acta Obstet Gynecol Scand 2001;80:830–4. (Level II-2)

Anthony S, Buitendijk SE, Zondervan KT, van Rijssel EJ, Verkerk PH. Episiotomies and the occurrence of severe perineal lacerations. Br J Obstet Gynaecol 1994;101:1064–7. (Level II-3)

Combs CA, Murphy EL, Laros RK Jr. Factors associated with postpartum hemorrhage with vaginal birth. Obstet Gynecol 1991;77:69–

76. (Level II-2)

Carroli G, Belizan J. Episiotomy for vaginal birth. The Cochrane Database of Systematic Reviews 1999, Issue 3. Art. No.: CD000081. DOI: 10.1002/14651858.CD000081. (Meta-Analysis)

1c.16 Quote verbatim, the specific guideline recommendation *(Including guideline # and/or page #)*:
Restricted use of episiotomy is preferable to routine use of episiotomy.

1c.17 Clinical Practice Guideline Citation: ACOG Practice Bulletin no. 71 April 2006

1c.18 National Guideline Clearinghouse or other URL: NA

1c.19 Grading of Strength of Guideline Recommendation. Has the recommendation been graded? Yes

1c.20 If guideline recommendation graded, identify the entity that graded the evidence including balance of representation and any disclosures regarding bias: ACOG

1c.21 System Used for Grading the Strength of Guideline Recommendation: GRADE

1c.22 If other, identify and describe the grading scale with definitions:

1c.23 Grade Assigned to the Recommendation: a

1c.24 Rationale for Using this Guideline Over Others: see above

Based on the NQF descriptions for rating the evidence, what was the developer's assessment of the quantity, quality, and consistency of the body of evidence?
1c.25 Quantity: High **1c.26** Quality: Moderate**1c.27** Consistency: High

1. Evidence, Performance Gap, Priority – Importance to Measure and Report

Extent to which the specific measure focus is evidence-based, important to making significant gains in healthcare quality, and improving health outcomes for a specific high-priority (high-impact) aspect of healthcare where there is variation in or overall less-than-optimal performance. ***Measures must be judged to meet all subcriteria to pass this criterion and be evaluated against the remaining criteria.***

1a. Evidence to Support the Measure Focus – See attached Evidence Submission Form
0470_Evidence_MSF5.0_Data.doc,0470_Evidence_MSF5.0_Data_Feb_2016.doc

1b. Performance Gap

Demonstration of quality problems and opportunity for improvement, i.e., data demonstrating:

- considerable variation, or overall less-than-optimal performance, in the quality of care across providers; and/or
- disparities in care across population groups.

1b.1. **Briefly explain the rationale for this measure** *(e.g., the benefits or improvements in quality envisioned by use of this measure)*
Episiotomy has been clearly linked with worse perineal tears and in turn its attendant complications. These are noted to include perineal pain, blood loss,and potential for wound break down/abscess formation and necrotizing fascitis. Predicated on these concerns, ACOG has called for "restricted use of episiotomy".

1b.2. **Provide performance scores on the measure as specified (current and over time) at the specified level of analysis.** *(This is required for endorsement maintenance. Include mean, std dev, min, max, interquartile range, scores by decile. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included). This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*
A systematic review comparing routine episiotomy with restrictive use reported that the groups varied between an overall incidence of 72.7% in the routine group versus 27.6% in the restricted-use group (ref Carrli).A validation exercise of this measure performed in 2010 by the National Perinatal Information Center,demonstrated that the rate had fallen to 16.2% with tremendous inter center variation(4.3% to 34.6%). This wide variation in this overuse measure suggest that there is tremendous opportunity to improve care for women through pubic reporting.
Period 2 shows a significant drop is Episiotomy rate (-7.8 percent change in unweighted average rate).
Period 3 analysis continues to show a significant drop in the unweighted average rate from 11.5% to 7.2% from CY 2010 to CY2014 on a 100% eligible cases across 68 hospitals.

1b.3. **If no or limited performance data on the measure as specified is reported in 1b2, then provide a summary of data from the literature that indicates opportunity for improvement or overall less than optimal performance on the specific focus of measurement.**
1.Carroli G, Belizan J. Episiotomy for vaginal birth. The Cochrane Database of Systematic Reviews 1999, Issue 3. Art. No.: CD000081. DOI: 10.1002/14651858.CD000081.
2. Internal data see validation exercise

1b.4. **Provide disparities data from the measure as specified (current and over time) by population group, e.g., by race/ethnicity, gender, age, insurance status, socioeconomic status, and/or disability.** *(This is required for endorsement maintenance. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities include.) This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*
Period 3 analysis continues to show a significant drop in the unweighted average rate from 11.5% to 7.2% from CY 2010 to CY2014 on a 100% eligible cases across 68 hospitals. The range of rates for CY 2014 was from a low of .8% to a high of 22.1% suggesting continued opportunities for improvement.

1b.5. **If no or limited data on disparities from the measure as specified is reported in 1b4, then provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement. Include citations.**
NA

1c. High Priority (previously referred to as High Impact)

The measure addresses:

- a specific national health goal/priority identified by DHHS or the National Priorities Partnership convened by NQF;

OR

- a demonstrated high-priority (high-impact) aspect of healthcare (e.g., affects large numbers of patients and/or has a substantial impact for a smaller population; leading cause of morbidity/mortality; high resource use (current and/or future); severity of illness; and severity of patient/societal consequences of poor quality).

1c.1. **Demonstrated high priority aspect of healthcare**
Affects large numbers, Frequently performed procedure
1c.2. **If Other:**

1c.3. **Provide epidemiologic or resource use data that demonstrates the measure addresses a high priority aspect of healthcare. List citations in 1c.4.**
Epidemiologic data has shown that episiotomy remains in high use despite the American Congress of Obstetrics & Gynecology recomendation limiting its use(ref ACOG). In 2,000, approximately 33% of vaginal births an episiotmy was used(ref Martin). A validation exercise of this measure performed in 2010 by the National Perinatal Information Center,demonstrated that the rate had fallen to 16.2% with tremendous inter center variation(4.3% to 34.6%). **Follow-up analysis through CY 2014 data shows a continuing variation (.8% to 22.1%).**

1c.4. **Citations for data demonstrating high priority provided in 1a.3**
1. ACOG- Practice Bulletin-"Episiotomy" No.71 2006
2.Martin JA, Hamilton BE, Ventura SJ, Menacker F, Park MM. Births: final data for 2000. Natl Vital Stat Rep 2002;50(5):1–101. (Level II-3)
3. Internal data see validation exercise

1c.5. **If a PRO-PM (e.g. HRQoL/functional status, symptom/burden, experience with care, health-related behaviors), provide evidence that the target population values the measured PRO and finds it meaningful.** *(Describe how and from whom their input was obtained.)*

2. Reliability and Validity—Scientific Acceptability of Measure Properties

Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. ***Measures must be judged to meet the subcriteria for both reliability and validity to pass this criterion and be evaluated against the remaining criteria.***

2a.1. Specifications The measure is well defined and precisely specified so it can be implemented consistently within and across organizations and allows for comparability. eMeasures should be specified in the Health Quality Measures Format (HQMF) and the Quality Data Model (QDM).

De.5. **Subject/Topic Area** *(check all the areas that apply):*
Perinatal and Reproductive Health, Perinatal and Reproductive Health : Perinatal

De.6. **Cross Cutting Areas** *(check all the areas that apply):*
Overuse

S.1. **Measure-specific Web Page** *(Provide a URL link to a web page specific for this measure that contains current detailed specifications including code lists, risk model details, and supplemental materials. Do not enter a URL linking to a home page or to general information.)*
Http://www.npic.org/NQF_Project/NQF_Episiotomy_Rate.php

S.2a. **If this is an eMeasure**, HQMF specifications must be attached. Attach the zipped output from the eMeasure authoring tool (MAT) - if the MAT was not used, contact staff. (Use the specification fields in this online form for the plain-language description of the specifications)
This is not an eMeasure **Attachment:**

S.2b. **Data Dictionary, Code Table, or Value Sets** *(and risk model codes and coefficients when applicable) must be attached. (Excel or csv file in the suggested format preferred - if not, contact staff)*
No data dictionary **Attachment:**

S.3. **For endorsement maintenance**, please briefly describe any changes to the measure specifications since last endorsement date and explain the reasons.
No changes except to update for use of ICD 10 codes beginning 10/1/2015

S.4. **Numerator Statement** *(Brief, narrative description of the measure focus or what is being measured about the target population, i.e., cases from the target population with the target process, condition, event, or outcome)*
IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
Number of episiotomy procedures (ICD-9 code 72.1, 72.21, 72.31, 72.71, 73.6; **ICD-10 PCS:0W8NXZZ** performed on women undergoing a vaginal delivery (excluding those with shoulder dystocia ICD-10; O66.0) during the analytic period- monthly,quarterly, yearly etc.

S.5. **Time Period for Data** *(What is the time period in which data will be aggregated for the measure, e.g., 12 mo, 3 years, look back to August for flu vaccination? Note if there are different time periods for the numerator and denominator.)*
Inpatient delivery stay.

S.6. **Numerator Details** *(All information required to identify and calculate the cases from the target population with the target process, condition, event, or outcome such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
IF an OUTCOME MEASURE, describe how the observed outcome is identified/counted. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
Any vaginal delivery with one of the ICD-9 codes for episiotomy- 72.1, 72.21, 72.31, 72.71, 73.6 **(ICD-10 PCS:0W8NXZZ)**

S.7. **Denominator Statement** *(Brief, narrative description of the target population being measured)*
All vaginal deliveries during the analytic period- monthly, quarterly, yearly etc. excluding those coded with a shoulder dystocia **ICD-**

10: O66.0).

S.8. **Target Population Category** *(Check all the populations for which the measure is specified and tested if any):*
Maternal Health

S.9. **Denominator Details** *(All information required to identify and calculate the target population/denominator such as definitions, specific data collection items/responses , code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
Any woman with a vaginal delivery calculated by either MS DRG 774,775,767,768

S.10. **Denominator Exclusions** *(Brief narrative description of exclusions from the target population)*
Women who have a coded complication of shoulder dystocia. In the case of shoulder dystocia, an episiotomy is performed to free the shoulder and prevent/mitigate birth injury to the infant.

S.11. **Denominator Exclusion Details** *(All information required to identify and calculate exclusions from the denominator such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
Vaginal deliveries coded with shoulder dystocia, ICD-9 code 660.41, 660.42**(ICD-10 CM : O66.0)**

S.12. **Stratification Details/Variables** *(All information required to stratify the measure results including the stratification variables, definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format with at S.2b)*
NA

S.13. **Risk Adjustment Type** (Select type. Provide specifications for risk stratification in S.12 and for statistical model in S.14-15)
No risk adjustment or risk stratification
If other:

S.14. **Identify the statistical risk model method and variables** *(Name the statistical method - e.g., logistic regression and list all the risk factor variables. Note - risk model development and testing should be addressed with measure testing under Scientific Acceptability)*
NA

S.15. **Detailed risk model specifications** *(must be in attached data dictionary/code list Excel or csv file. Also indicate if available at measure-specific URL identified in S.1.)*
Note: Risk model details (including coefficients, equations, codes with descriptors, definitions), should be provided on a separate worksheet in the suggested format in the Excel or csv file with data dictionary/code lists at S.2b.
URL

S.15a. **Detailed risk model specifications** (*if not provided in excel or csv file at S.2b*)

S.16. **Type of score:**
Rate/proportion
If other:

S.17. **Interpretation of Score** *(Classifies interpretation of score according to whether better quality is associated with a higher score, a lower score, a score falling within a defined interval, or a passing score)*
Better quality = Lower score

S.18. **Calculation Algorithm/Measure Logic** *(Describe the calculation of the measure score as an ordered sequence of steps including identifying the target population; exclusions; cases meeting the target process, condition, event, or outcome; aggregating data; risk adjustment; etc.)*
A. Identify all vaginal deliveries for time period in question
B. Exclude those coded with shoulder dystocia to obtain denominator cases

C. Of the denominator cases, identify those coded with an episiotomy
D Divide numerator by denominator and calculate the rate or convert a percent

S.19. **Calculation Algorithm/Measure Logic Diagram URL or Attachment** *(You also may provide a diagram of the Calculation Algorithm/Measure Logic described above at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
No diagram provided

S.20. **Sampling** *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*
IF a PRO-PM, identify whether (and how) proxy responses are allowed.
NA

S.21. **Survey/Patient-reported data** *(If measure is based on a survey, provide instructions for conducting the survey and guidance on minimum response rate.)*
IF a PRO-PM, specify calculation of response rates to be reported with performance measure results.

S.22. **Missing data** (specify how missing data are handled, e.g., imputation, delete case.)
Required for Composites and PRO-PMs.

S.23. **Data Source** *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*
If other, please describe in S.24.
Administrative claims, Paper Medical Records

S.24. **Data Source or Collection Instrument** *(Identify the specific data source/data collection instrument e.g. name of database, clinical registry, collection instrument, etc.)*
IF a PRO-PM, identify the specific PROM(s); and standard methods, modes, and languages of administration.
UB04 claims data.

S.25. **Data Source or Collection Instrument** *(available at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
No data collection instrument provided

S.26. **Level of Analysis** *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*
Facility

S.27. **Care Setting** *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*
Hospital/Acute Care Facility
If other:

S.28. **COMPOSITE Performance Measure** - Additional Specifications *(Use this section as needed for aggregation and weighting rules, or calculation of individual performance measures if not individually endorsed.)*

2a. Reliability – See attached Measure Testing Submission Form
2b. Validity – See attached Measure Testing Submission Form
0470_MeasureTesting_MSF5.0_Data.zip,**0470_2016_Measure_Testing_MSF_6.5.pdf**

2. RELIABALITY & VALIDITY – SCIENTIFIC ACCEPTABILITY OF MEASURE PROPERTIES

Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. **(evaluation criteria)**

Measure testing must demonstrate adequate reliability and validity in order to be recommended for endorsement. Testing may be conducted for data elements and/or the computed measure score. Testing information and results should be entered in the appropriate field. Supplemental materials may be referenced or attached in item 2.1. See guidance on measure testing.

2a2. Reliability Testing. (*Reliability testing was conducted with appropriate method, scope, and adequate demonstration of reliability*.)

2a2.1 Data/Sample *(Description of the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)***:**

Testing was performed by analyzing data from 12 NPIC/QAS member hospitals for Period 1 : 10/1/08-9/30/09 and Period 2: 10/1/09-3/31/10. For Period 1 totalled 66,306 eligible deliveries and 9,626 episiotomy cases, Period 2: 31,496 eligigble deliveries and 4,259 episiotomy cases.

2a2.2 Analytic Method *(Describe method of reliability testing & rationale)***:**

Testing was performed by analyzing data from 12 NPIC/QAS member hospitals for Period 1 : 10/1/08-9/30/09 and Period 2: 10/1/09-3/31/10. For Period 1 totalled 66,306 eligible deliveries and 9,626 episiotomy cases, Period 2: 31,496 eligigble deliveries and 4,259 episiotomy cases.

2a2.3 Testing Results *(Reliability statistics, assessment of adequacy in the context of norms for the test conducted)***:**

For Period 1, 7 of 9 responding hospitals (63.4%) confirmed the coding on the sample episiotomy cases matched exactly with the medical record. One hospital had a discrepancy of 1 case and the second hospital did not indicate the degree of discrepancy. 8 of 9 (89%) indicated they felt the administrative data set was a consistent and reliable source of episiotomy data. For Period 2, 11 hospitals responded; 4 of the 11 (36.6%) found all cases, with and without episiotomies to be correctly coded. The remaining 7 found 1-4 cases with codes not matching documentation, evenly split between those with and without episiotomies.

2b. VALIDITY. Validity, Testing, including all Threats to Validity: H M L I

2b1.1 Describe how the measure specifications *(measure focus, target population, and exclusions)* **are consistent with the evidence cited in support of the measure focus (***criterion 1c)* **and identify any differences from the evidence:**

Testing was performed by analyzing data from 12 NPIC/QAS member hospitals for Period 1 : 10/1/08-9/30/09 and Period 2: 10/1/09-3/31/10. For Period 1 totalled 66,306 eligible deliveries and 9,626 episiotomy cases, Period 2: 31,496 eligigble deliveries and 4,259 episiotomy cases.

2b2. Validity Testing. (*Validity testing was conducted with appropriate method, scope, and adequate demonstration of validity.)*

2b2.1 Data/Sample *(Description of the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)***:**

In addition to auditing their sample of cases, hosptials were asked to rate the reliability, validity, feasibility and usability of this measure

2b2.2 Analytic Method *(Describe method of validity testing and rationale; if face validity, describe systematic assessment)***:** NQF #0470 Incidence of Episiotomy, Last Updated Date: Jan 02, 2013

In addition to auditing their sample of cases, hosptials were asked to rate the reliability, validity, feasibility and usability of this measure

2b2.3 Testing Results *(Statistical results, assessment of adequacy in the context of norms for the test conducted; if face validity, describe results of systematic assessment)***:**

In period 2, 9 of 11 hospitals (81.8%) indicated they felt episiotomy rate is a valid measure of the quality of care at a hospital; the other 2 felt the measure was valid but should be looked at at the provider level since providers will determine whether to perform an episiotomy or not

POTENTIAL THREATS TO VALIDITY. (*All potential threats to validity were appropriately tested with adequate results.*)

2b3. Measure Exclusions. (*Exclusions were supported by the clinical evidence in 1c or appropriately tested with results demonstrating the need to specify them.*)

2b3.1 **Data/Sample for analysis of exclusions** *(Description of the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)***:**
In the case of shoulder dystocia, an episiotomy is performed to free the shoulder and prevent/mitigate birth injury to the infant. This based on expert opinion nonetheless the incidence of this is estimated to be 1%

2b3.2 **Analytic Method** (*Describe type of analysis and rationale for examining exclusions, including exclusion related to patient preference)***:** NA

2b3.3 **Results** *(Provide statistical results for analysis of exclusions, e.g., frequency, variability, sensitivity analyses):* NA

2b4. Risk Adjustment Strategy. (*For outcome measures, adjustment for differences in case mix (severity) across measured entities was appropriately tested with adequate results.*)

2b4.1 **Data/Sample** *(Description of the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)***:** NA

2b4.2 **Analytic Method (***Describe methods and rationale for development and testing of risk model or risk stratification including selection of factors/variables)***:** NA

2b4.3 **Testing Results** *(Statistical risk model: Provide quantitative assessment of relative contribution of model risk factors; risk model performance metrics including cross-validation discrimination and calibration statistics, calibration curve and risk decile plot, and assessment of adequacy in the context of norms for risk models. Risk stratification: Provide quantitative assessment of relationship of risk factors to the outcome and differences in outcomes among the strata)***:** NA
2b4.4 **If outcome or resource use measure is not risk adjusted, provide rationale and analyses to justify lack of adjustment:**
NA

2b5. Identification of Meaningful Differences in Performance. (*The performance measure scores were appropriately analyzed and discriminated meaningful differences in quality.*)

2b5.1 **Data/Sample** *(Describe the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)***:**

Data was obtained from 12 National Perinatal Information Center/Quality Analytic Services(NPIC/QAS) member hospitals that are a subset of our larger membership. The data are from their submitted discharge abstract data submitted quarterly to NPIC/QAS as part of an ongoing quality evaluation process. There were two periods of time analyzed: 10/1/08-9/30/09 and 10/1/09-3/31/10

2b5.2 **Analytic Method** *(Describe methods and rationale to identify statistically significant and practically/meaningfully differences in performance)***:**

T-Test for changes in the episiotomy rate and laceration rate for the 12 hospitals between Time1 and Time2 show a significant drop in the episiotomy rate; the laceration rate also dropped but not significantly.
Pearson function shows a significant inverse correlation between decreasing episiotomy rate and laceration rate in Time2.

2b5.3 **Results** *(Provide measure performance results/scores, e.g., distribution by quartile, mean, median, SD, etc.; identification of statistically significant and meaningfully differences in performance)*:
T-Test for changes in the episiotomy rate and laceration rate for the 12 hospitals between Time1 and Time2 show a significant drop in the episiotomy rate; the laceration rate also dropped but not significantly.
Pearson function shows a significant inverse correlation between decreasing episiotomy rate and laceration rate in Time2.

2b6. Comparability of Multiple Data Sources/Methods. (*If specified for more than one data source, the various approaches result in comparable scores.*)

2b6.1 **Data/Sample** *(Describe the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)*: NA

2b6.2 **Analytic Method** *(Describe methods and rationale for testing comparability of scores produced by the different data sources specified in the measure)*: NA
2b6.3 **Testing Results** *(Provide statistical results, e.g., correlation statistics, comparison of rankings; assessment of adequacy in the context of norms for the test conducted)*: NA

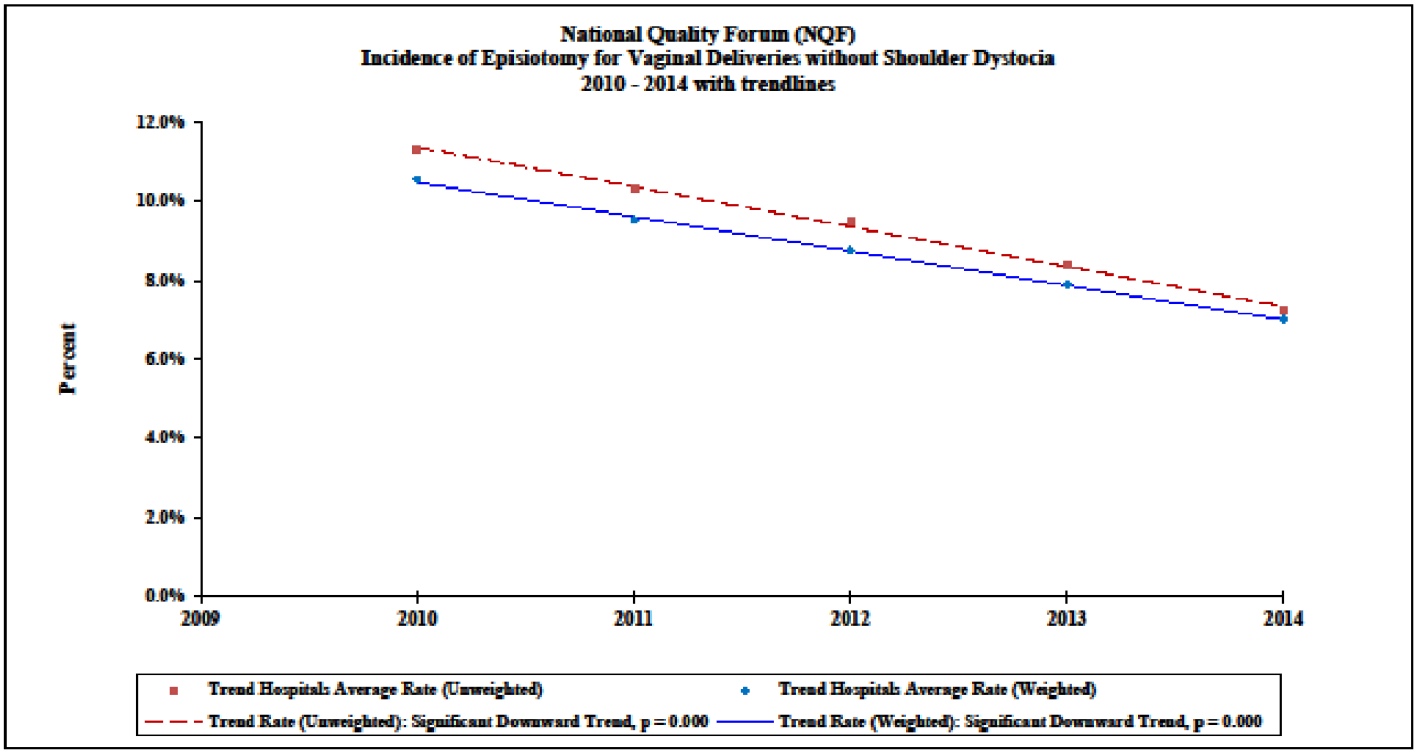
2c. Disparities in Care: H M L I NA (*If applicable, the measure specifications allow identification of disparities.*)
2c.1 **If measure is stratified for disparities, provide stratified results** *(Scores by stratified categories/cohorts)*: NA
2c.2 **If disparities have been reported/identified (e.g., in 1b), but measure is not specified to detect disparities, please explain:** NA
2.1-2.3 **Supplemental Testing Methodology Information:**

Attachment
Episiotomy validation results.xlsx

National Quality Forum (NQF)
Incidence of Episiotomy for Vaginal Deliveries without Shoulder Dystocia
2010 - 2014

n = 63 hospitals in NPIC/QAS Trend database	2010	2011	2012	2013	2014	Pct Change 2010 - 2014
Total vaginal deliveries without shoulder dystocia	168,949	169,729	169,042	167,941	169,181	0.1%
Vaginal deliveries without shoulder dystocia with episiotomy	17,808	16,153	14,803	13,231	11,862	-33.4%
Percent (Trend Hospitals weighted average) *	10.5%	9.5%	8.8%	7.9%	7.0%	-33.5%
Percent (Trend Hospitals unweighted average) *	11.3%	10.3%	9.5%	8.4%	7.3%	-35.9%
Range						
Low	0.6%	0.7%	0.6%	0.8%	0.8%	
High	35.1%	32.8%	29.3%	25.2%	22.1%	

*** Trend rate (Weighted and Unweighted): Significant Downward Trend, p = 0.000**



	2010	2011	2012	2013	2014	Pct change
Trend Rate (unweighted)	11.3%	10.3%	9.5%	8.4%	7.3%	-35.9%
Trend Rate (weighted)	10.3%	9.3%	8.8%	7.9%	7.0%	-33.3%
Trend Numerator (weighted)	17,808	16,153	14,803	13,231	11,862	
Trend Denominator (weighted)	168,949	169,729	169,042	167,941	169,181	

3. Feasibility

Extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

3a. Byproduct of Care Processes

For clinical measures, the required data elements are routinely generated and used during care delivery (e.g., blood pressure, lab test, diagnosis, medication order).

3a.1. **Data Elements Generated as Byproduct of Care Processes.**

Coded by someone other than person obtaining original information (e.g., DRG, ICD-9 codes on claims)

If other:

3b. Electronic Sources

The required data elements are available in electronic health records or other electronic sources. If the required data are not in electronic health records or existing electronic sources, a credible, near-term path to electronic collection is specified.

3b.1. **To what extent are the specified data elements available electronically in defined fields?** *(i.e., data elements that are needed to compute the performance measure score are in defined, computer-readable fields)*

ALL data elements are in defined fields in electronic claims

3b.2. **If ALL the data elements needed to compute the performance measure score are not from electronic sources, specify a credible, near-term path to electronic capture, OR provide a rationale for using other than electronic sources.**

3b.3. **If this is an eMeasure, provide a summary of the feasibility assessment in an attached file or make available at a measure-specific URL.**

Attachment:

3c. Data Collection Strategy

Demonstration that the data collection strategy (e.g., source, timing, frequency, sampling, patient confidentiality, costs associated with fees/licensing of proprietary measures) can be implemented (e.g., already in operational use, or testing demonstrates that it is ready to put into operational use). For eMeasures, a feasibility assessment addresses the data elements and measure logic and demonstrates the eMeasure can be implemented or feasibility concerns can be adequately addressed.

3c.1. **Describe what you have learned/modified as a result of testing and/or operational use of the measure regarding data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, time and cost of data collection, other feasibility/implementation issues.**

IF a PRO-PM, consider implications for both individuals providing PROM data (patients, service recipients, respondents) and those whose performance is being measured.

See above.

3c.2. **Describe any fees, licensing, or other requirements to use any aspect of the measure as specified** *(e.g., value/code set, risk model, programming code, algorithm)*.

4. Usability and Use

Extent to which potential audiences (e.g., consumers, purchasers, providers, policy makers) are using or could use performance results for both accountability and performance improvement to achieve the goal of high-quality, efficient healthcare for individuals or populations.

4a. Accountability and Transparency

Performance results are used in at least one accountability application within three years after initial endorsement and are publicly reported within six years after initial endorsement (or the data on performance results are available). If not in use at the time of initial endorsement, then a credible plan for implementation within the specified timeframes is provided.

4.1. **Current and Planned Use**

NQF-endorsed measures are expected to be used in at least one accountability application within 3 years and publicly reported within 6 years of initial endorsement in addition to performance improvement.

Planned	Current Use (for current use provide URL)
Public Reporting Quality Improvement with Benchmarking (external benchmarking to multiple organizations) Quality Improvement (Internal to the specific organization)	

4a.1. **For each CURRENT use, checked above, provide:**

- Name of program and sponsor
- Purpose
- Geographic area and number and percentage of accountable entities and patients included

National Perinatal Information Center, Inc. (NPIC) Metric reported to member and military hospitals for quality review and bench marking. Total 400,000+ deliveries across all 130 hospitals and the majority of states.

4a.2. **If not currently publicly reported OR used in at least one other accountability application (e.g., payment program, certification, licensing) what are the reasons?** (*e.g., Do policies or actions of the developer/steward or accountable entities restrict access to performance results or impede implementation?*)

4a.3. **If not currently publicly reported OR used in at least one other accountability application, provide a credible plan for implementation within the expected timeframes -- any accountability application within 3 years and publicly reported within 6 years of initial endorsement.** (*Credible plan includes the specific program, purpose, intended audience, and timeline for implementing the measure within the specified timeframes. A plan for accountability applications addresses mechanisms for data aggregation and reporting.*)

4b. Improvement

Progress toward achieving the goal of high-quality, efficient healthcare for individuals or populations is demonstrated. If not in use for performance improvement at the time of initial endorsement, then a credible rationale describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.

4b.1. **Progress on Improvement. (Not required for initial endorsement unless available.)**

Performance results on this measure (current and over time) should be provided in 1b.2 and 1b.4. Discuss:

- **Progress (trends in performance results, number and percentage of people receiving high-quality healthcare)**
- **Geographic area and number and percentage of accountable entities and patients included**

The NPIC Trend data base from 2010- 2014 shows a significant drop in episiotomy rates using both weight and unweighted averages. The unweighted average (preferred method for hospital to hospital comparisons) went from 11.3% in 2010 to 7.3% in 2014. The range in 2014 was .8% to 22.1 % indicating continued opportunities for improvement.

4b.2. If no improvement was demonstrated, what are the reasons? If not in use for performance improvement at the time of initial endorsement, provide a credible rationale that describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.
NA

4c. Unintended Consequences

The benefits of the performance measure in facilitating progress toward achieving high-quality, efficient healthcare for individuals or populations outweigh evidence of unintended negative consequences to individuals or populations (if such evidence exists).

4c.1. Were any unintended negative consequences to individuals or populations identified during testing; OR has evidence of unintended negative consequences to individuals or populations been reported since implementation? If so, identify the negative unintended consequences and describe how benefits outweigh them or actions taken to mitigate them.
Prior to Q4, 2008 ICD-9 coding updates, the repair of a laceration could be coded with clear indication that the laceration was the result of a tear of an episiotomy. This sample of hospitals sucessfully petition CMS to update the codes to include 73.6 Episiotomy, allowing hospitals to clearly identify the episiotomy and the repair procedure. We assume this coding convention has been adopted and therefore the susceptability to inaccuracies, errors and unintended consequences is small. For our sample, in Period 1, the 9 responding hospitals that re-abstracted a 5% sample of their episiotomy cases found a very high degree of match between the administrative data and abstracted data. 7 of 9 had an exact match; 1 hospital had a 1 case descrepancy and the second hospital said the discrepancy was small but did not identify the count. In Period 2, 4 of the 11 hospitals had no discrepancy in their coding of cases with or without an episiotomy. The 7 hospitals with errors the count of errors was from 1 to 4 cases split between cases with and without epiotomies. The overall rate of coding error was on less 3%.

5. Comparison to Related or Competing Measures

If a measure meets the above criteria and there are endorsed or new related measures (either the same measure focus or the same target population) or competing measures (both the same measure focus and the same target population), the measures are compared to address harmonization and/or selection of the best measure.

5. Relation to Other NQF-endorsed Measures
Are there related measures (conceptually, either same measure focus or target population) or competing measures (conceptually both the same measure focus and same target population)? If yes, list the NQF # and title of all related and/or competing measures.
No

5.1a. **List of related or competing measures (selected from NQF-endorsed measures)**

5.1b. **If related or competing measures are not NQF endorsed please indicate measure title and steward.**

5a. Harmonization

The measure specifications are harmonized with related measures;
OR
The differences in specifications are justified

5a.1. **If this measure conceptually addresses EITHER the same measure focus OR the same target population as NQF-endorsed measure(s):**
Are the measure specifications completely harmonized?
Yes

5a.2. **If the measure specifications are not completely harmonized, identify the differences, rationale, and impact on interpretability and data collection burden.**

5b. Competing Measures

The measure is superior to competing measures (e.g., is a more valid or efficient way to measure);
OR
Multiple measures are justified.

5b.1. **If this measure conceptually addresses both the same measure focus and the same target population as NQF-endorsed measure(s):**
Describe why this measure is superior to competing measures (e.g., a more valid or efficient way to measure quality); OR provide a rationale for the additive value of endorsing an additional measure. (Provide analyses when possible.)

Appendix

A.1 Supplemental materials may be provided in an appendix. All supplemental materials (such as data collection instrument or methodology reports) should be organized in one file with a table of contents or bookmarks. If material pertains to a specific submission form number, that should be indicated. Requested information should be provided in the submission form and required attachments. There is no guarantee that supplemental materials will be reviewed.

Attachment:

Contact Information

Co.1 **Measure Steward (Intellectual Property Owner):** Christiana Care Health System
Co.2 **Point of Contact:** Matthew, Hoffman, mhoffman@christianacare.org, 302-733-6610-
Co.3 **Measure Developer if different from Measure Steward:** National Perinatal Information Center

Co.4 **Point of Contact:** Matthew, Hoffman, mhoffman@christianacare.org

Additional Information

Ad.1 **Workgroup/Expert Panel involved in measure development**
Provide a list of sponsoring organizations and workgroup/panel members' names and organizations. Describe the members' role in measure development.

Measure Developer/Steward Updates and Ongoing Maintenance
Ad.2 **Year the measure was first released:** 2011
Ad.3 **Month and Year of most recent revision:** 10, 2015
Ad.4 **What is your frequency for review/update of this measure?** As needed
Ad.5 **When is the next scheduled review/update for this measure?** 10, 2016

Ad.6 **Copyright statement:**
Ad.7 **Disclaimers:**

Ad.8 **Additional Information/Comments:**

MEASURE WORKSHEET

This document summarizes the evaluation of the measure as it progresses through NQF's Consensus Development Process (CDP). The information submitted by measure developers/stewards is included after the Brief Measure Information, Preliminary Analysis, and Pre-meeting Public and Member Comments sections.

To navigate the links in the worksheet: Ctrl + click link to go to the link; ALT + LEFT ARROW to return

Brief Measure Information

NQF #: 0471
Measure Title: PC-02 Cesarean Birth
Measure Steward: The Joint Commission
Brief Description of Measure: This measure assesses the number of nulliparous women with a term, singleton baby in a vertex position delivered by cesarean birth. This measure is part of a set of five nationally implemented measures that address perinatal care (PC-01: Elective Delivery, PC-03: Antenatal Steroids, PC-04: Health Care-Associated Bloodstream Infections in Newborns, PC-05: Exclusive Breast Milk Feeding).
Developer Rationale: A reduction in the number of nulliparous patients with live term singleton newborns in vertex position (NTSV) delivering by cesarean birth will result in increased patient safety, a substantial decrease in maternal and neonatal morbidity and substantial savings in health care costs. Successful quality improvement efforts incorporate audit and feedback strategies combined with provider and nurse education, guidelines and peer review

The measure will assist health care organizations (HCOs) to track nulliparous patients with live term singleton newborns in vertex position delivering by cesarean birth to reduce the occurrence. Nulliparous women have 4-6 times the cesarean birth rate than multiparous women thus the NTSV population is the largest driver of primary cesarean birth rate. Furthermore nulliparity varies greatly among hospitals (20% to 60%) making it the most important risk factor for stratification or adjustment. NTSV has the large variation among facilities thus identifying an important population on which to focus quality improvement efforts.

In addition, a reduction in primary cesarean births will reduce the number of women having repeat cesarean births (currently >90% of mothers who have a primary cesarean birth will have a Cesarean for all her subsequent births). Thus, improvement in the rates of cesarean birth for the first birth will reduce the morbidity of all future births and avoid all the controversies with trial of labor after cesarean/elective repeat cesareans.

Numerator Statement: The outcome being measured is: Patients with cesarean births with ICD-10-PCS Principal Procedure Code or ICD-10-PCS Other Procedure Codes for cesarean birth as defined in Appendix A, Table 11.06 available at:
http://manual.jointcommission.org/releases/ TJC2016A ~~TJC2015B2~~/
Denominator Statement: The outcome target population being measured is: Nulliparous patients delivered of a live term singleton newborn in vertex presentation ICD-10-PCS Principal or Other Diagnosis Codes for delivery as defined in Appendix A, Tables 11.01.1 available at:
http://manual.jointcommission.org/releases/ TJC2016A ~~TJC2015B2~~/
Denominator Exclusions: • ICD-10-CM Principal Diagnosis Code or ICD-10-CM Other Diagnosis Codes for multiple gestations and other presentations as defined in Appendix A, Table 11.09
• Less than 8 years of age
• Greater than or equal to 65 years of age
• Length of Stay >120 days
• ~~Enrolled in clinical trials~~
• Gestational Age < 37 weeks or UTD

Measure Type: Outcome
Data Source: Paper Medical Records
Level of Analysis: Facility, Population : National

IF Endorsement Maintenance – Original Endorsement Date: Oct 24, 2008 **Most Recent Endorsement Date:** Mar 30, 2012

Maintenance of Endorsement -- Preliminary Analysis

To maintain NQF endorsement endorsed measures are evaluated periodically to ensure that the measures still meets the NQF endorsement criteria ("maintenance"). The emphasis for maintaining endorsement is focused on how effective the measure is for promoting improvements in quality. Endorsed measures should have some experience from the field to inform the evaluation. The emphasis for maintaining endorsement is noted for each criterion.

Criteria 1: Importance to Measure and Report

1a. Evidence
Maintenance measures – less emphasis on evidence unless there is new information or change in evidence since the prior evaluation.

1a. Evidence. The evidence requirements for a health outcomes measure include providing rationale that supports the relationship of the health outcome to processes or structures of care. The guidance for evaluating the clinical evidence asks if the relationship between the measured health outcome and at least one clinical action is identified and supported by the stated rationale.

Summary of evidence from 2012 evaluation:

- The Committee noted that " ACOG says this is the "optimal measure" for Cesarean section because it focuses on first-time, uncomplicated pregnancy. The measure looks at the outcome of labor management. "
- The developer reports that "Among primary cesarean deliveries, more subjective indications (non-reassuring fetal status and arrest of dilation) contributed larger proportions than more objective indications (malpresentation, maternal-fetal, and obstetric conditions)."
- Cesarean sections are associated with increased risk of obstetric hemorrhage, uterine infection, and increased costs to the health care system.

Changes to evidence from last review

☒ **The developer attests that there have been no changes in the evidence since the measure was last evaluated.**
☐ **The developer provided updated evidence for this measure.**

Exception to evidence NA

Questions for the Committee:

- *The developer attests the underlying evidence for the measure has not changed since the last NQF endorsement review. Does the Committee agree the evidence basis for the measure has not changed and there is no need for repeat discussion and voting on Evidence?*

Preliminary rating for evidence: ☒ **Pass** ☐ **No Pass**

1b. Gap in Care/Opportunity for Improvement and 1b. disparities
Maintenance measures – increased emphasis on gap and variation

1b. Performance Gap. The performance gap requirements include demonstrating quality problems and opportunity for improvement.

- The Healthy People 2020 goal for cesarean section is 23.9%. In 2010, when The Joint Commission began reporting

on this measure, approximately 165 hospitals reported the data with an average measure rate of 26.7%.

- In 2014, The Joint Commission began mandating reporting of the measure for all accredited hospitals with 1100 births or more annually.

The developers provided the following data (unadjusted rates):

	2011	2012	2013	2014
# hospitals	166	169	200	1388
# patients	33,3797	33,944	44,679	363,400
National aggregate rate	26.3%	26.2%	25.9%	26.8%
10^{th}-25^{th} -90^{th} %tile	16.8 -20.8– 35.5%	17.1– 20.2--36.1%	16.4- 20- 35.7%	17.6 - 21.2 -36.3 %
Mean hospital rate (SD)	0.26636 (0.09659)	0.26335 (0.08582)	0.25792 (0.09181)	0.26732(0.09064)

Disparities

- The developer reports on literature that indicates the following factors are associated with greater odds of cesarean section: African American race, marital status, patient type, insurance type, and age older than 35 years.
- No data using this measure is presented with stratifications for these characteristics.

Questions for the Committee:

- *Is there a gap in care that warrants a national performance measure?*

Preliminary rating for opportunity for improvement: ☒ **High** ☐ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 1: Importance to Measure and Report (including 1a, 1b, 1c)

1a. Evidence to Support Measure Focus

Comments:

The evidence is directly related to the outcome of this measure (outcome measure).

1b. Performance Gap

Comments:

Performance data was provided. Large variation in performance by facility, practice level and provider. Data on the differences in population subgroups is provided. There are large discrepancies in care.

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability

2a1. Reliability Specifications

Maintenance measures – no change in emphasis – specifications should be evaluated the same as with new measures

2a1. Specifications requires the measure, as specified, to produce consistent (reliable) and credible (valid) results about the quality of care when implemented.

Data source(s): Paper medical records; Vital Records reports, delivery logs and clinical information systems were added as acceptable data sources to help hospitals identify all cases with =>37 weeks gestation

Specifications:

- The specifications have been updated to ICD-10.
- The initial patient population is now identified with ICD-10-PCS-Principal or Other Procedure Codes for delivery instead of diagnosis codes for pregnancy, since the ICD-10-CM Principal or Other Diagnosis Codes do not indicate whether the delivery took place during the hospitalization.

- Cases with a gestational age of UTD were added to the denominator excluded populations, since UTD is highly correlated with no prenatal care.
- Two data elements are used for the observed outcome and to calculate the numerator, and eight data elements are used to identify the target population and calculate the denominator.
- The ~~measure is stratified by patient age~~developer has removed the age stratification adjustment..
- A calculation algorithm is included.
- Sampling is allowed.
- ~~The adjustment using direct standardization for age is unchanged. More information below.~~

Questions for the Committee:

- *Are the changes to the numerator and denominator appropriate?*
- *Are all the data elements clearly defined? Are all appropriate codes included?*
- *Is the logic or calculation algorithm clear?*
- *Is it likely this measure is consistently implemented?*

2a2. Reliability Testing Testing attachment
Maintenance measures – less emphasis if no new testing data provided

2a2. Reliability testing demonstrates if the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise enough to distinguish differences in performance across providers.

For maintenance measures, summarize the reliability testing from the prior review (2012):

- The measure was tested using Inter-rater Reliability (IRR) by the ORYX vendor, for 108 hospitals with 13,279 records. IRR is an appropriate method of assessing data element reliability for chart abstraction. The agreement rate for the data element "Gestational age" was 89.75% and the data element "Parity" was 97.43%.

Describe any updates to testing No updates to reliability testing

SUMMARY OF TESTING

Reliability testing level ☐ **Measure score** ☒ **Data element** ☐ **Both**

Reliability testing performed with the data source and level of analysis indicated for this measure ☒ **Yes** ☐ **No**

Method(s) of reliability testing Inter-rater reliability testing (above)

Results of reliability testing The results of the reliability testing only provided % agreement. No statistical test such as Kappa or ICC was provided.

Guidance from the Reliability Algorithm

Precise specifications (Box 1) → Empirical reliability testing (Box 2) → Data element testing (Box 8) → appropriate method- IRR (Box 9) → high or moderate confidence of reliability of numerator data element → moderate (highest rating possible)

Questions for the Committee:

- *The developer has not provided any additional testing for reliability since the last NQF endorsement review. Does the Committee agree there is no need for repeat discussion and voting on reliability?*

Preliminary rating for reliability: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

2b. Validity
Maintenance measures – less emphasis if no new testing data provided

2b1. Validity: Specifications

2b1. Validity Specifications. This section should determine if the measure specifications are consistent with the evidence.

Specifications consistent with evidence in 1a. ☒ **Yes** ☐ **Somewhat** ☐ **No**
Specification not completely consistent with evidence

Question for the Committee:
- *Are the specifications consistent with the evidence?*

2b2. Validity testing

2b2. Validity Testing should demonstrate the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality.

For maintenance measures, summarize the validity testing from the prior review: Face validity only.
Continued face validity has been determined through feedback from measure users.

Describe any updates to validity testing
Additional validity testing of the measure score was performed.

SUMMARY OF TESTING
Validity testing level ☒ **Measure score** ☐ **Data element testing against a gold standard** ☐ **Both**

Method of validity testing of the measure score:
- ☐ **Face validity only**
- ☒ **Empirical validity testing of the measure score**

Validity testing method:
- Data from Q3 and Q4, 2015, and Q1 and Q2, 2015, (1,345 hospitals submitted 2,695,467 inpatient records) for all the Joint Commission perinatal care measures. Measure convergent validity for PC-02 was assessed using hospitals patient level data from The Joint commission warehouse. The observed rates were used for PC-02 in the analysis in this section -hospital rates were not age standardized.
- Spearman rank order correlation was performed to correlate the performance across all perinatal care measures.

Validity testing results:
- The developer provides scatter plots and a correlation table. They note that the correlation on the performance between this measure and others in the perinatal care set (other than PC-04: Health Care-Associated Bloodstream Infections in Newborns), although in the expected direction and statistically significant, are relatively weak.

Questions for the Committee:
- *Is the test sample adequate to generalize for widespread implementation?*
- *Do the results demonstrate sufficient validity so that conclusions about quality can be made?*
- *Do you agree that the score from this measure as specified is an indicator of quality?*

2b3-2b7. Threats to Validity

2b3. Exclusions:

This measure has six exclusions:
- less than 8 years or 65 or greater years of age, which are included in the initial patient population exclusion.
- *ICD-9-CM Principal Diagnosis Code or ICD-9-CM Other Diagnosis Codes* for multiple gestations and other

presentations as defined in Appendix A, Table 11.09; this is to exclude cases from the denominator.
- Patients who have a length of stay greater than 120 days, to harmonize with other CMS/Joint Commission measures; this is included in the initial patient population exclusion.
- Patients enrolled in a clinical trial: overall occurrence of n=729, or 0.06%
- Patients with Gestational Age < 37 weeks or UTD (which generally indicates no prenatal care): n=113,520 or 9.7%.

The developer notes that "The difference between the 10th and 90th percentiles of the distribution of exclusion rates is narrow indicating that the occurrence is random and likely would not bias performance results."

Questions for the Committee:

- *Are the exclusions consistent with the evidence?*
- *Are any patients or patient groups inappropriately excluded from the measure?*
- *Are the exclusions/exceptions of sufficient frequency and variation across providers to be needed (and outweigh the data collection burden)?*

2b4. Risk adjustment: **Risk-adjustment method** ☒~~☐~~ **None** ☐ **Statistical model** ☐ **Stratification** ☐~~☒~~ **Other**

~~**Conceptual rationale for SDS factors included? ☒ Yes ☐ No**~~

~~**SDS factors included in risk model? ☐ Yes ☒ No**~~

~~**Risk adjustment summary**~~

- ~~The cesarean birth measure in the Perinatal Care measure set (PC-02) uses the direct standardization method to risk-adjust each organization's observed measure rate. For this measure, one applies the national maternal distribution age at first delivery to an organization's aggregated measure population, by weighting the observed cesarean birth rates for each age group according to their national frequency. They are then summed to give the adjusted rate. The adjusted (or expected) rate is then interpreted as what the cesarean birth rate would be expected to be if the organization performed at the national rate for each age group.~~
- ~~Direct standardization can only adjust for one factor. However, the developer fit a model with age category and the SDS risk factors (African American race, other non-white race and Hispanic ethnicity) to evaluate the impact of the SDS risk factors.~~
- ~~The developer states that African American and Hispanic ethnicity were statistically significant, but "the overall discrimination of the model did not change much from the model with just the age categories (c=0.581 for the model with age categories alone versus c=0.595 for the model with age categories and ethnicity)."~~

Questions for the Committee:

- ~~*Is the risk adjustment strategy included in the measure appropriate?*~~
- ~~*Are the candidate and final variables included in the risk adjustment model adequately described for the measure to be implemented?*~~
- ~~*Do you agree with the developer's rationale that there is no conceptual basis for adjusting this measure for SDS factors?*~~
- ~~*Do you agree with the developer's decision, based on their analysis, to not include SDS factors in their risk adjustment model?*~~*removal of the age stratification appropriate?*

2b5. Meaningful difference *(can statistically significant and clinically/practically meaningful differences in performance measure scores can be identified)*:

The developer reports the following descriptive statistics for PC-02 measure:
N=1,345 hospitals n = 1,169,924

Min = 0%
Mean: 26.2%

Percentile 10%: 18%
Percentile 25%: 21%
Median: 25.4%
Percentile 75%: 30.4%
Percentile 90%: 36%
Max = 100%

The Joint Commission's Target Analysis uses two methods: Given the national average for a performance measure, a target range is constructed. Using generalized linear mixed models methodology (also known as hierarchical models), a predicted estimate of an HCO's performance, with a corresponding 95% confidence interval, is generated. This confidence interval is compared to the target range, to determine the HCOs' rating. The estimate of the organization's true performance is based on both the data from that organization and on data from the entire set of reporting organizations. The object of target analysis is to compare a health care organizations (HCO) data against a comparative norm for the purpose of evaluating performance improvement opportunities. When an organization's performance level is statistically significantly different from a comparative norm, it is considered a statistical deviation. A statistical deviation may be desirable or undesirable depending on the "direction of improvement" of the measure.

PC-02 Distribution of Outliers
2011 1st Quarter Data:
Scores on this measure: N=160, Mean 26.7%, SD 0.12953
10th Percentile= 14%
25th Percentile= 19.4%
50th Percentile=26%
75th Percentile= 32.5%
90th Percentile= 40%
159 (100%) Neutral – results not significantly different from target range

Question for the Committee:

- *Does this measure identify meaningful differences about quality?*

2b6. Comparability of data sources/methods:
N/A

2b7. Missing Data
The developer reports "Any file with missing data will result in a measure category assignment of X and rejection of the file from the warehouse, unless the data are not used to process the measure. If the data are used to process the measure and have been reported by the abstractor as No/UTD, the case will result in a measure category assignment of E (i.e., failed measure, not rejected)."

Guidance from the algorithm:
Specifications aligned with evidence (Box 1) → Threats to validity assessed (Box 2) →empirical validity testing (Box 3) → Testing of measure score (Box 6) → appropriate method (Box 7) →moderate certainty (Box 8b) → moderate

Preliminary rating for validity: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 2: Scientific Acceptability of Measure Properties (including all 2a, 2b, and 2d)

2a1. & 2b1. Specifications
Comments:
Clearly defined. Measure should also have the following exclusions for c/s: previa, HIV+, and myomectomy. Algorithm is clear. No concerns about consistent implementation.
There are additional contraindications for c/s delivery that should be added to the exclusion list for this measure: Previa, HIV+, and previous myomectomy.

2a2. Reliability Testing
Comments:

Adequate scope. Appropriate method. Inter-rater reliability occurred providing %agreement but did not have statistical testing.

2b2. Validity Testing

Comments:

Adequate sample with Spearman correlation. Variation in rates across facilities shows much work to do in this area nationally.

2b3. Exclusions Analysis

2b4. Risk Adjustment/Stratification for Outcome or Resource Use Measures

2b5. Identification of Statistically Significant & Meaningful Differences In Performance

2b6. Comparability of Performance Scores When More Than One Set of Specifications

2b7. Missing Data Analysis and Minimizing Bias

Comments:

No due to the relatively small number of cases that will have missing data.

Criterion 3. Feasibility

Maintenance measures – no change in emphasis – implementation issues may be more prominent

3. Feasibility is the extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

- According to the developer, "hospitals using this performance measure generally collect measure data via manual review of the paper medical record, the EMR or a combination of both. Collected data are submitted to The Joint Commission on a quarterly basis, by way of contracted performance measurement system vendors".
- The specifications are freely available and there are no fees or licensing requirements.
- The developer is currently retooling the measure for capture from electronic sources and plans to test in 2016.

Questions for the Committee:

- *What is the burden/costs for hospitals in collecting data for this measure?*
- *Are the required data elements routinely generated and used during care delivery?*
- *Are the required data elements available in electronic form, e.g., EHR or other electronic sources?*

Preliminary rating for feasibility: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 3: Feasibility

3a. Byproduct of Care Processes

3b. Electronic Sources

3c. Data Collection Strategy

Comments:

Routinely generated data elements. Should be available electronically. Time consuming for accurate data collection. Deep dives are required into patient records for accurate data abstraction.

Criterion 4: Usability and Use

Maintenance measures – increased emphasis – much greater focus on measure use and usefulness, including both impact /improvement and unintended consequences

4. Usability and Use evaluate the extent to which audiences (e.g., consumers, purchasers, providers, policymakers) use or could use performance results for both accountability and performance improvement activities.

Current uses of the measure

Publicly reported? ☒ **Yes** ☐ **No**

Current use in an accountability program? ☒ **Yes** ☐ **No**
OR
Planned use in an accountability program? ☐ **Yes** ☐ **No**

Accountability program details
The measure is in use in the Joint Commission's Hospital Accreditation Program, is publically reported in the Joint Commission's Annual Report, *America's Hospitals: Improving Quality and Safety,* and is used for internal quality improvement via the Perinatal Care Certification program run by the Joint Commission. 3,300 hospitals are accredited by the Joint Commission and 10 hospitals are included in the Perinatal Care Certification.

Improvement results
The developers report that the rate has remained relatively the same: 26.7% reported from 165 hospitals in 2010 and 26.8% reported from 1388 hospitals for 2014.
However, improvements were seen in the lower quartile (21.1%) and the 10th percentile (17.6%) hospitals.
The Healthy People 2020 goal is 23.9%.

Unexpected findings (positive or negative) during implementation
The developers reported on two unexpected consequences of the measure:

- Patients who didn't receive prenatal care were inappropriately included, but the gestational age data could not be determined. The developer has removed patients with gestational age unable to be determined.
- Because of sampling, some hospitals had small denominator populations and reported higher rates. The developer added three data sources (vital records reports, delivery logs, and clinical information systems) to help hospitals identify all appropriate cases to include the denominator population size.

Potential harms
None reported

Feedback:

***Questions for the Committee**:*

- *Committee members should share any experience with use of this measure. Are you aware of any unintended consequences or unexpected findings from use of the measure?*
- *What does experience with this measure tell various stakeholders?*
- *How can the performance results be used to further the goal of high-quality, efficient healthcare?*
- *Do the benefits of the measure outweigh any potential unintended consequences?*

Preliminary rating for usability and use: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 4: Usability and Use

4a. Accountability and Transparency
4b. Improvement
4c. Unintended Consequences
Comments:
TJC Perinatal Core Measure. Need a consistent agreed upon goal nationally. Healthy People 2020 states a goal of 23.9% but this doesn't specify nulliparous. It specifies women with no prior cesarean birth. There are unintended consequences seen in some facilities with rates <11% with increased chorio, etc. Nationally we have so much work to do in this area.

Criterion 5: Related and Competing Measures
Related or competing measures • 2892 Birthrisk Cesarean Birth Measure (newly submitted for 2016) **Harmonization** • To be discussed

Pre-meeting public and member comments

- I have concerns about measure #0471 as well as the recent actions of the measure steward. My concerns stem from the fact that the Joint Commission (JC) has failed to disclose that they are aware of a fatal error in measure #0471: PC-02 Cesarean Birth. I have been attempting to inform the JC since April 2010 that the direct standardization age risk adjustment used in measure PC-02 contains a fatal flaw [1]. This error was not immediately obvious to others including the authors, the editors, the JC and the National Quality Forum (NQF). I was finally able to get the JC to understand the fatal flaw in July 2015 which they acknowledged in an email this past September. They informed me that instead of recalling the flawed measure that they would just be dropping the risk adjustment.
  I was horrified by their decision since dropping the risk adjustment from measure PC-02 creates a new and significantly more flawed cesarean birth measure which has never been tested, validated or endorsed. The JC's website indeed confirms that they have dropped the risk adjustment from measure PC-02 v2016A [2]. Surprisingly, their website clearly indicates that this new cesarean birth measure is "NQF-ENDORSED". The actions of the JC are especially concerning because these actions make it very confusing as to which cesarean birth measure was vetted by the NQF and recently adopted by The Core Quality Measure Collaborative.
  I understand the significant problem that I have created by exposing a fatal flaw in the widely distributed measure #0471. However, the decision by the JC to conceal the flaw from the NQF will only make the problem worse. The longer it takes to recall a fatally flawed measure the more significant the problem will become.
  It was clear to me six years ago that the fatal flaw in PC-02 would eventually require a recall of the measure but unfortunately my concerns were ignored by the JC. Ignoring my concerns in 2010 has resulted in six wasted years in the effort to accurately measure cesarean birth utilization. Therefore, it would be extremely irresponsible of me if I didn't alert everyone involved of the current actions of the JC before they waste another six years and potentially adversely affect not only the millions of women who are giving birth each year but also the hard working healthcare personnel that care for them.
  I understand the extremely serious nature of my concerns and stand ready to provide any and all evidence required in support of my concerns. My motives are clear and my conviction is unwavering. Women who give birth deserve better.
  References:
  1. https://www.birthrisk.com/Public/FatalFlaw.pdf
  2. https://manual.jointcommission.org/releases/TJC2016A/MIF0167.html

- Developer Response:
  The Joint Commission has had numerous, detailed communications with the commenter on this subject, and is of the opinion that current evidence contradicts his contentions. The final decision to remove all risk-adjustment from this measure was made after submitting the measure to NQF and is based on evidence from two recent studies [1,2] which have shown that hospitals with a high maternal age population also have a low body mass index (BMI) and conversely, those with low maternal age have a high BMI. When tested against a more robust risk adjusted model (age, BMI, race, hypertension, diabetes), the studies found differences limited to 1-2%. Because age and BMI tend to cancel each other out in the risk-adjustment models in these studies, and because BMI often cannot be calculated because height is often not recorded in the medical record, the

Joint Commission's Perinatal Care Technical Advisory Panel has recommended using the simple cesarean birth rates without further risk adjustment. Therefore, effective with discharges beginning July 1, 2016, The Joint Commission has removed all risk adjustments until such time as data are available demonstrating the need for risk adjustment and the feasibility of collecting any risk factors required. The Joint Commission also notes that the commenter seems to be principally concerned with promoting his own proprietary and competing "measure" rather than with substantive issues with this measure.

[1] Caceres IA, Arcaya M, Declercq E, Belanoff CM, Janakiraman V, et al. (2013) Hospital Differences in Cesarean Deliveries in Massachusetts (US) 2004–2006:The Case against Case-Mix Artifact. PLoS ONE 8(3): e57817. doi:10.1371/journal.pone.0057817

[2] Main E. (2014) Nuliparous, Term, Singleton, Vertex (NTSV) Cesarean Birth Rates: extreme hospital variation is not changed by adjustment for case-mix. Oral Presentation: Pacific Coast Obstetrics and Gynecology Society

1. IMPACT, OPPORTUITY, EVIDENCE - IMPORTANCE TO MEASURE AND REPORT

Importance to Measure and Report is a threshold criterion that must be met in order to recommend a measure for endorsement. All three subcriteria must be met to pass this criterion. See guidance on evidence.

Measures must be judged to be important to measure and report in order to be evaluated against the remaining criteria. (evaluation criteria)

1c.1 Structure-Process-Outcome Relationship *(Briefly state the measure focus, e.g., health outcome, intermediate clinical outcome, process, structure; then identify the appropriate links, e.g., structure-process-health outcome; process- health outcome; intermediate clinical outcome-health outcome)*:
The focus of the measure is to decrease the number of cesarean sections >> population determined >> population assessed >> patient delivers vaginally >> improved maternal and fetal outcomes >> decreased length of stay and fetal morbidity and mortality.

1c.2-3 Type of Evidence *(Check all that apply)*:
Clinical Practice Guideline
Systematic review of body of evidence (other than within guideline development)

1c.4 Directness of Evidence to the Specified Measure *(State the central topic, population, and outcomes addressed in the body of evidence and identify any differences from the measure focus and measure target population)*:
The central topic for the measure is the reduction in the number of nulliparous patients with live term singleton newborns in vertex position delivering by cesarean section. The target population for the performance measure is consistent with the body of evidence supporting quality improvement strategies to reduce the number of NTSV cesarean sections.

1c.5 Quantity of Studies in the Body of Evidence *(Total number of studies, not articles)*: The body of literature looking at cesarean section rates is very large with over 5,000 articles published since 1980. Over 1,000 articles have focused on the quality issues around nulliparous (sometimes called primiparous) cesarean section rates. Specifically, the low-risk first-birth cesarean rate has been examined by over 250 retrospective cohort and prospective observational studies. Synonyms in the literature include: NTSV (nulliparous, term, singleton, vertex) cesarean rate, Standard primip cesarean rate, and the Robson 10-category cesarean classification system (of which NTSV is the key driver). Throughout these studies, NTSV cesarean has emerged as the group with highest variation and greatest contribution to the rise in cesarean rates both in the US and internationally.

1c.6 Quality of Body of Evidence (*Summarize the certainty or confidence in the estimates of benefits and harms to patients across studies in the body of evidence resulting from study factors. Please address: a) study design/flaws; b) directness/indirectness of the evidence to this measure (e.g., interventions, comparisons, outcomes assessed, population included in the evidence); and c) imprecision/wide confidence intervals due to few patients or events):* The quality of evidence supporting the reduction in the number of NTSV cesarean sections is quite high. The rate of severe obstetric hemorrhage has significantly increased (by 50%) over the last 15 years in the U.S. There has also been a 270% increase in blood transfusions, with both hemorrhage and transfusions correlated to the rise in cesarean deliveries. Infection is the most common serious complication of cesarean delivery with typical rates of 3 to 9%. As noted, ACOG has evaluated cesarean sections and made a recommendation to adopt the NTSV cesarean section rate as a national metric to address through quality improvement interventions. Studies of quality improvement initiatives aimed at reducing NTSV cesarean sections have also noted a decrease in the number of such deliveries as well as a subsequent decrease in the number of maternal and neonatal morbidities.

1c.7 Consistency of Results across Studies *(Summarize the consistency of the magnitude and direction of the effect):* The body of evidence consistently supports the reduction of NTSV cesarean deliveries. Studies looking at multi faceted quality improvement interventions also show a decrease in the number of NTSV cesarean sections. NTSV cesarean delivery rates show much more consistency than total or primary cesarean delivery rates as it is much more tightly focused on labor management issues.

1c.8 Net Benefit *(Provide estimates of effect for benefit/outcome; identify harms addressed and estimates of effect; and net benefit - benefit over harms)*:
As described before, nulliparous patients with live term singleton newborns in vertex position delivering vaginally result in improved maternal and neonatal outcomes and will result in substantial savings in health care costs. Furthermore, the benefit is extended to all future pregnancies—if the first birth is a cesarean, then 90% of the remainder will be cesareans with health risks markedly increasing for each additional cesarean.

1c.9 Grading of Strength/Quality of the Body of Evidence. Has the body of evidence been graded? No

1c.10 If body of evidence graded, identify the entity that graded the evidence including balance of representation and any disclosures regarding bias: Not Applicable

1c.11 System Used for Grading the Body of Evidence: Other

1c.12 If other, identify and describe the grading scale with definitions: Although grading of the evidence was not determined during our systematic review, it was determined that the guideline developers accounted for a balanced representation of information, and provided information that was accessible and met the requirements set out in this measure maintenance form.

1c.13 Grade Assigned to the Body of Evidence: Not Applicable

1c.14 Summary of Controversy/Contradictory Evidence: There is no documented evidence regarding controversy related to the reduction of NTSV cesarean sections. A review of recent studies also supports the use of quality improvement interventions to further reduce the number of such deliveries.

1c.15 Citations for Evidence other than Guidelines*(Guidelines addressed below)*:

- Barber EL, Lundsberg LS, Belanger K, Pettker CM, Funai EF, Illuzzi JL. (2011). Indications contributing to the increasing cesarean delivery rate. Am J Obstet Gynecol. 118(1):29-38.
- Brennan, DJ, Robson, MS, Murphy, M, O´Herlihy, C. (2009). Comparative analysis of international cesarean delivery rates using 10-group classification identifies significant variation in spontaneous labor. American Journal of Obstetrics and Gynecology. 201(3):308 e301-308.
- Coonrod, DV, Drachman, D, Hobson, P, Manriquez, M. (2008). Nulliparous term singleton vertex cesarean delivery rates: institutional and individual level predictors. American Journal of Obstetrics and Gynecology. 198(6):694 e691-611; discussion 694 e611.
- Ehrenthal, DB, Jiang, X, & Strobino, DM. (2010). Labor induction and the risk of a cesarean delivery among nulliparous women at term. Am J Obstet Gynecol . 116(1):35-42.
- Getahun D, Strickland D, Lawrence JM, Fassett MJ, Koebnick C, Jacobsen SJ. (2009). Racial and ethnic disparities in the trends in primary cesarean delivery based on indications. Am J Obstet Gynecol. 201(4):422 e421-427.
- Main, EK, Moore, D, Farrell, B, et al. (2006). Is there a useful cesarean birth measure? Assessment of the nulliparous term singleton vertex cesarean birth rate as a tool for obstetric quality improvement. American Journal of Obstetrics and Gynecology. 194(6):1644-1651; discussion 1651-1642.
- Main, EK. (1999). Reducing cesarean birth rates with data-driven quality improvement activities. Pediatrics.103(1 Suppl E):374-383.
- US Department of Health and Human Services (DHHS). (2000). Healthy People 2010. Washington, DC. Retrieved on Setember 26, 2011 at: http://www.healthypeople.gov/2010
- US Department of Health and Human Services (DHHS). (2010). Healthy People 2020. Washington, DC. Retrieved on Setember 26, 2011 at: http://www.healthypeople.gov/2020

1c.16 Quote verbatim, <u>the specific guideline recommendation</u> *(Including guideline # and/or page #)*:

The American College of Obstetricians and Gynecologists (ACOG) in their monograph on evaluating Cesarean delivery rates, recommended this measure for benchmarking cesarean section rates on page 35:

"Institutions and practitioners should consider reviewing their cesarean delivery rates with these benchmarks for 1) nulliparous women with term singleton fetuses with vertex presentations and, 2) multiparous women with one previous low-transverse cesarean delivery and term singleton fetuses with vertex presentations." II-3

1c.17 Clinical Practice Guideline Citation:

- American Collge of Obstetricians and Gynecologists (ACOG) Task Force on Cesarean Delivery Rates. Evaluation of Cesarean Delivery. 2000. Washington, DC.

1c.18 National Guideline Clearinghouse or other URL:

1c.19 Grading of Strength of Guideline Recommendation. Has the recommendation been graded? Yes

1c.20 If guideline recommendation graded, identify the entity that graded the evidence including balance of representation and any disclosures regarding bias: The American College of Obstetrics and Gynecology

1c.21 System Used for Grading the Strength of Guideline Recommendation: USPSTF

1c.22 If other, identify and describe the grading scale with definitions:

1c.23 Grade Assigned to the Recommendation: Level II

1c.24 Rationale for Using this Guideline Over Others: The American College of Obstetricians and Gynecologists (ACOG) is the nation´s leading group of professionals providing health care for women. The monograph developed by the ACOG Task Force on Cesarean Delivery provides obstetricians and gynecologists with current information on established techniques and clinical management guidelines. The American College of Obstetricians and Gynecologists (the College) continuously surveys the field for advances to be incorporated in these series and monitors existing bulletins to ensure they are current. Individual bulletins are withdrawn from and added to the series on a continuing basis and reaffirmed periodically.

Specifically, NTSV cesarean section rate is preferred over total or primary cesarean rates as it more narrowly focuses on the population at greatest risk (nulliparous women in labor) with the greatest long-term consequences and

Based on the NQF descriptions for rating the evidence, what was the <u>developer's assessment</u> of the quantity, quality, and consistency of the body of evidence?

1c.25 Quantity: High **1c.26** Quality: High**1c.27** Consistency**:** High

1. Evidence, Performance Gap, Priority – Importance to Measure and Report

Extent to which the specific measure focus is evidence-based, important to making significant gains in healthcare quality, and improving health outcomes for a specific high-priority (high-impact) aspect of healthcare where there is variation in or overall less-than-optimal performance. ***Measures must be judged to meet all subcriteria to pass this criterion and be evaluated against the remaining criteria.***

1a. Evidence to Support the Measure Focus – See attached Evidence Submission Form
0471_Evidence_MSF5.0_Data-635908032652533791.doc

1b. Performance Gap

Demonstration of quality problems and opportunity for improvement, i.e., data demonstrating:

- considerable variation, or overall less-than-optimal performance, in the quality of care across providers; and/or
- disparities in care across population groups.

1b.1. **Briefly explain the rationale for this measure** *(e.g., the benefits or improvements in quality envisioned by use of this measure)*
A reduction in the number of nulliparous patients with live term singleton newborns in vertex position (NTSV) delivering by cesarean birth will result in increased patient safety, a substantial decrease in maternal and neonatal morbidity and substantial savings in health care costs. Successful quality improvement efforts incorporate audit and feedback strategies combined with provider and nurse education, guidelines and peer review

The measure will assist health care organizations (HCOs) to track nulliparous patients with live term singleton newborns in vertex position delivering by cesarean birth to reduce the occurrence. Nulliparous women have 4-6 times the cesarean birth rate than multiparous women thus the NTSV population is the largest driver of primary cesarean birth rate. Furthermore nulliparity varies greatly among hospitals (20% to 60%) making it the most important risk factor for stratification or adjustment. NTSV has the large variation among facilities thus identifying an important population on which to focus quality improvement efforts.

In addition, a reduction in primary cesarean births will reduce the number of women having repeat cesarean births (currently >90% of mothers who have a primary cesarean birth will have a Cesarean for all her subsequent births). Thus, improvement in the rates of cesarean birth for the first birth will reduce the morbidity of all future births and avoid all the controversies with trial of labor after cesarean/elective repeat cesareans.

1b.2. **Provide performance scores on the measure as specified (current and over time) at the specified level of analysis.** *(This is required for endorsement maintenance. Include mean, std dev, min, max, interquartile range, scores by decile. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included). This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*
Nulliparous term singleton vertex cesarean births continue to remain above The Healthy People 2020 goal of 23.9% (DHHS, 2010). The Perinatal Care (PC) core measures were added as a new core measure set in 2010 for hospitals to select in order to meet their ORYX performance measurement requirement for Joint Commission accreditation purposes. At that time, approximately 165 hospitals reported the data with an average measure rate of 26.7% (n=25,143 patients). In January 2014, The Joint Commission required mandatory reporting of the PC measure set for all accredited hospitals with 1100 births or more annually. 1388 hospitals reported the data with an average rate of 26.8% (n=363,400 patients). The 2014 performance gap persists with improvement noted primarily in the lower quartile (21.1%) and 10th percentile (17.6%) hospitals. It is important to note that a performance gap of 12.4% exists for the 90th percentile of hospitals performing at 36.3% (if 23.9% is considered goal performance) The 2014 mean rate of 26.7% also remains above the HP 2020 goal. The threshold for mandatory reporting was recently lowered to 300 births annually effective January 2016. The new reporting requirement will now capture approximately 80% of all accredited birthing hospitals. As a result, the rates may increase with the addition of approximately 821 more hospitals reporting data. Below is the specified level of analysis for PC-02 beginning with discharges April 1, 2010 through December 31, 2014.
Note: PC-02 hospital rates listed in this section were not age standardized.

2Q 2010: 25,143 denominator cases; 6,708 numerator cases; 165 hospitals; 26.7% national aggregate rate; 0.26636 mean of hospital rates; 0.09659 standard deviation; 40.0% 90th percentile rate; 31.9% 75th percentile rate/upper quartile; 26.3% 50th percentile rate/median rate; 20.5% 25th percentile rate/lower quartile; and 15.4% 10th percentile rate.
CY 2011: 33,379 denominator cases; 8,779 numerator cases; 166 hospitals; 26.3% national aggregate rate; 0.26283 mean of hospital rates; 0.08961 standard deviation; 35.3% 90th percentile rate; 30.7% 75th percentile rate/upper quartile; 25.7% 50th percentile rate/median rate; 20.8% 25th percentile rate/lower quartile; and 16.8% 10th percentile rate.
CY 2012: 33,944 denominator cases; 9,428 numerator cases; 169 hospitals; 26.2% national aggregate rate; 0.26335 mean of hospital

rates; 0.08582 standard deviation; 36.1% 90th percentile rate; 31.1% 75th percentile rate/upper quartile; 25% 50th percentile rate/median rate; 20.2% 25th percentile rate/lower quartile; and 17.1% 10th percentile rate.
CY 2013: 44,679 denominator cases; 11,553 numerator cases; 200 hospitals; 25.9% national aggregate rate; 0.25792 mean of hospital rates; 0.09181 standard deviation; 35.7% 90th percentile rate; 30.8%% 75th percentile rate/upper quartile; 25% 50th percentile rate/median rate; 20% 25th percentile rate/lower quartile; and 16.4% 10th percentile rate.
CY 2014: 363,400 denominator cases; 97,270 numerator cases; 1388 hospitals; 26.8% national aggregate rate; 0.26732 mean of hospital rates; 0.09064 standard deviation; 36.3% 90th percentile rate; 31.0% 75th percentile rate/upper quartile; 25.9% 50th percentile rate/median rate; 21.2% 25th percentile rate/lower quartile; and 17.6% 10th percentile rate.

1b.3. **If no or limited performance data on the measure as specified is reported in 1b2, then provide a summary of data from the literature that indicates opportunity for improvement or overall less than optimal performance on the specific focus of measurement.**

- US Department of Health and Human Services (DHHS). (2010). Healthy People 2020. Washington, DC. Retrieved on November 3, 2015 at: http://www.healthypeople.gov/2020

1b.4. **Provide disparities data from the measure as specified (current and over time) by population group, e.g., by race/ethnicity, gender, age, insurance status, socioeconomic status, and/or disability.** *(This is required for endorsement maintenance. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities include.) This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*
Within California, Braveman et al. (1995) examined whether there was an independent association with cesarean delivery and women's socioeconomic characteristics and type of hospital. There was large variation between the types of hospital ownership. County hospitals had 47% fewer primary cesarean deliveries than noted in private for-profit hospitals. After controlling for insurance, personal, community, medical and hospital characteristics, African American women were 24% more likely to undergo cesarean delivery than whites. 2007 data from the California Office of Statewide Hospital Planning and Development (OSHPD) continued to report that African American populations tend to have higher NTSV cesarean rates even after adjustment for clinical factors, again leading to suspicion that provider attributes may play a role (OSHPD, 2007).

A large retrospective cohort study of cesarean deliveries at a large regional US hospital was conducted from 2003-2006. Factors associated with greater odds of cesarean delivery included African American race, marital status, patient type, insurance type, and age older than 35 years (Ehrenthal et al., 2010). Zhang, et al. (2013) also noted when compared to other racial groups, African American women were younger, more likely to have a cesarean delivery, have longer lengths of stay and increased Medicaid costs. Despite having the same insurance coverage, African American women were also more likely to have adverse pregnancy outcomes when compared to White and Hispanic women.

1b.5. **If no or limited data on disparities from the measure as specified is reported in 1b4, then provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement. Include citations.**

- Braveman, P, Egerter, S, Edmonston, F, Verdon, M. (1995). Racial/ethnic differences in the likelihood of cesarean delivery, California. Am J Public Health. 85(5):625-630.
- California Office of Statewide Hospital Planning and Development (OSHPD). (2007). Utilization Rates for Selected Medical Procedures in California Hospitals, Retrieved from the Internet on September 26, 2011 at: http://www.oshpd.ca.gov/HID/Products/PatDischargeData/ResearchReports/HospIPQualInd/Vol-Util_IndicatorsRpt/2007Util.pdf
- Ehrenthal, DB, Jiang, X, & Strobino, DM. (2010). Labor induction and the risk of a cesarean delivery among nulliparous women at term. Am J Obstet Gynecol . 116(1):35-42.
- Zhang, S., Cardarelli, K., Shim, R., Ye, J., Booker, K. & Rust, G. (2013). Racial disparities in economic and clinical outcomes of pregnancy among Medicaid recipients. Matern Child Health J. 17(8):1518-25

1c. High Priority (previously referred to as High Impact)

The measure addresses:

- a specific national health goal/priority identified by DHHS or the National Priorities Partnership convened by NQF; OR
- a demonstrated high-priority (high-impact) aspect of healthcare (e.g., affects large numbers of patients and/or has a substantial impact for a smaller population; leading cause of morbidity/mortality; high resource use (current and/or future); severity of illness; and severity of patient/societal consequences of poor quality).

1c.1. **Demonstrated high priority aspect of healthcare**
Affects large numbers, A leading cause of morbidity/mortality, Frequently performed procedure, High resource use
1c.2. **If Other:**

1c.3. Provide epidemiologic or resource use data that demonstrates the measure addresses a high priority aspect of healthcare. List citations in 1c.4.

The growing support for the claim that provider-dependent indications are contributing to the overall increase among cesareans can be seen from the results of two recent studies examining the drivers for the increase in cesarean deliveries. Barber et al. (2011) at Yale analyzed primary and repeat cesareans from 2003 to 2009. Among primary cesarean deliveries, more subjective indications (non-reassuring fetal status and arrest of dilation) contributed larger proportions than more objective indications (malpresentation, maternal-fetal, and obstetric conditions). Similarly, Getahun et al. (2009) examined the causes for the rise in cesarean deliveries among different racial and ethnic groups in Kaiser Permanente Southern California over the last 17 years. Their findings were similar to those from Yale. In a retrospective cohort study conducted by Ehrenthal et al. (2010), labor induction was associated with a twofold increase in the odds of a cesarean delivery after adjustment for confounders. This was more pronounced among a low-risk group of women without major complications.

Beyond the medical burden to mothers and babies, the financial burden on payers is large: facility charges for cesarean are nearly twice that for vaginal delivery ($24,700 vs. $14,500). In California alone, the additional heath care costs to the system are conservatively estimated to be over $300 million annually (Main et al., 2011)

The most frequent causes of severe maternal morbidity are obstetric hemorrhage (bleeding) and uterine infection. These are significantly more common with cesarean surgery and also represent the two leading causes of hospital readmission in the first 30 days post delivery. A recent CDC analysis showed that the rate of severe obstetric hemorrhage has significantly increased (by 50%) over the last 15 years in the U.S. There has also been a 270% increase in blood transfusions, with both hemorrhage and transfusions correlated to the rise in cesarean deliveries. Infection is the most common serious complication of cesarean delivery with typical rates of 3 to 9% (Kuklina et al., 2009).

The American College of Obstetrics and Gynecology (ACOG) report, “Evaluation of Cesarean Delivery,” recognizes the importance of the Nulliparous, Term Singleton Vertex (NTSV) population as the optimal focus for measurement and quality improvement action. Furthermore, the report identified a target of 15.5% for NTSV births, one recommended by the National Center for Health Statistics. Although the ACOG target rate was directed at the NTSV cesarean delivery rate, the recommendation has been widely misread as recommending a 15.5% total cesarean delivery rate (ACOG, 2000).

In its 2000 report, ACOG formally recommended that NTSV Cesarean Delivery Rate be used to benchmark all U.S. hospitals and practitioners. This measure and target was then endorsed by the United States Healthy People 2010 objectives: 16-9 (DHHS, 2000). This same measure has been reaffirmed in Healthy People 2020 (MICH-7.1) but with a more modest target of a 23.9% NTSV rate (DHHS, 2010).

1c.4. Citations for data demonstrating high priority provided in 1a.3

• American College of Obstetricians and Gynecologists (ACOG). (2000). Task Force on Cesarean Delivery Rates. Evaluation of Cesarean Delivery.
• American College of Obstetricians and Gynecologists (ACOG). (2014). Obstetric Care Consensus No. 1, Safe Prevention of the Primary Cesarean Delivery.
• Barber EL, Lundsberg LS, Belanger K, Pettker CM, Funai EF, Illuzzi JL. (2011). Indications contributing to the increasing cesarean delivery rate. Am J Obstet Gynecol. 118(1):29-38.
• Ehrenthal, DB, Jiang, X, & Strobino, DM. (2010). Labor induction and the risk of a cesarean delivery among nulliparous women at term. Am J Obstet Gynecol. 116(1):35-42.
• Getahun D, Strickland D, Lawrence JM, Fassett MJ, Koebnick C, Jacobsen SJ. (2009). Racial and ethnic disparities in the trends in primary cesarean delivery based on indications. Am J Obstet Gynecol. 201(4):422 e421-427.
• Kuklina EV, Meikle SF, Jamieson DJ, et al. (2009).Severe Obstetric Morbidity in the United States: 1998-2005. Obstetrics and Gynecology.113(2):293-299.
• Main EK, Morton CH, Hopkins D, Giuliani G, Melsop K and Gould JB. (2011). Cesarean Deliveries, Maternal Outcomes, and Opportunities for Change in California. Palo Alto, CA: CMQCC. Available at: www.cmqcc.org
• Main, E., Morton, C,, Melsop, K., Hopkins, D., Giuliani, G. & Gould, J. (2012). Creating a public agenda for maternity safety and quality in cesarean delivery. Obstet Gynecol. 120(5):1194-8.
• Osterman, M. & Martin, J. (2014). Trends in low-risk cesarean delivery in the United States, 1990-2013. Natl Vital Stat Rep. 5;63(6):1-16.
• Simpson, K. (2014). Labor management evidence update: potential to minimize risk of cesarean birth in healthy women. J Perinat Neonatal Nurs. 28(2):108-16
• US Department of Health and Human Services (DHHS). (2000). Healthy People 2010. Washington, DC. Retrieved on September 26,

2011 at: http://www.healthypeople.gov/2010
• US Department of Health and Human Services (DHHS). (2010). Healthy People 2020. Washington, DC. Retrieved on November 3, 2015 at: http://www.healthypeople.gov/2020

1c.5. **If a PRO-PM (e.g. HRQoL/functional status, symptom/burden, experience with care, health-related behaviors), provide evidence that the target population values the measured PRO and finds it meaningful.** *(Describe how and from whom their input was obtained.)*
Not Applicable

2. Reliability and Validity—Scientific Acceptability of Measure Properties

Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. ***Measures must be judged to meet the subcriteria for both reliability and validity to pass this criterion and be evaluated against the remaining criteria.***

2a.1. Specifications The measure is well defined and precisely specified so it can be implemented consistently within and across organizations and allows for comparability. eMeasures should be specified in the Health Quality Measures Format (HQMF) and the Quality Data Model (QDM).

De.5. **Subject/Topic Area** *(check all the areas that apply):*
Perinatal and Reproductive Health, Perinatal and Reproductive Health : Perinatal

De.6. **Cross Cutting Areas** *(check all the areas that apply):*
Disparities, Overuse, Safety : Complications

S.1. **Measure-specific Web Page** *(Provide a URL link to a web page specific for this measure that contains current detailed specifications including code lists, risk model details, and supplemental materials. Do not enter a URL linking to a home page or to general information.)*
http://manual.jointcommission.org/releases/ TJC2016A ~~TJC2015B2~~/

S.2a. **If this is an eMeasure**, HQMF specifications must be attached. Attach the zipped output from the eMeasure authoring tool (MAT) - if the MAT was not used, contact staff. (Use the specification fields in this online form for the plain-language description of the specifications)
This is not an eMeasure **Attachment:**

S.2b. **Data Dictionary, Code Table, or Value Sets** *(and risk model codes and coefficients when applicable) must be attached. (Excel or csv file in the suggested format preferred - if not, contact staff)*
Attachment **Attachment:** PC02_ICD_~~and_CS_Direct_Standardization_~~Template_Nulliparous_Births.xlsx

S.3. **For endorsement maintenance**, please briefly describe any changes to the measure specifications since last endorsement date and explain the reasons.

- The measure name was changed from Cesarean Section to Cesarean Birth throughout the measure specifications in order to conform to the American College of Obstetricians & Gynecologists (ACOG) national standardized description for this type of delivery.
- All ICD-9-CM diagnosis codes and ICD-9-CM procedure codes were converted to ICD-10-CM diagnosis and ICD-10-PCS procedure codes throughout the measure specifications. The Department of Health and Human Services (HHS) mandated that all entities covered by the Health Insurance Portability and Accountability Act (HIPAA) must transition to a new set of codes for electronic health care transactions on October 1, 2015.
- The initial patient population is now identified with ICD-10-PCS-Principal or Other Procedure Codes for delivery instead of diagnosis codes for pregnancy, since the ICD-10-CM Principal or Other Diagnosis Codes do not indicate whether the delivery took place during the hospitalization.
- Cases with a gestational age of UTD were added to the denominator excluded populations, since UTD is highly correlated with no prenatal care.
- The name of table 11.09 was changed from Contraindications to Vaginal Delivery to Multiple Gestations and Other Presentations in order to clarify that the intent of the table was only to exclude cases with multiple gestations and presentations other than vertex from the measure.
- Vital Records reports, delivery logs and clinical information systems were added as acceptable data sources to help hospitals identify all cases with =>37 weeks gestation, so that 100% of these cases could be reviewed to increase the denominator population size and to identify the number of previous live births to reduce the burden of data abstraction.
- The Parity data element was re-named to Number of Previous Live Births to reflect only previous live births in order to enable the use of Vital Records reports for hospitals to obtain this information electronically.

S.4. **Numerator Statement** *(Brief, narrative description of the measure focus or what is being measured about the target population, i.e., cases from the target population with the target process, condition, event, or outcome)*
IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
The outcome being measured is: Patients with cesarean births with ICD-10-PCS Principal Procedure Code or ICD-10-PCS Other

Procedure Codes for cesarean birth as defined in Appendix A, Table 11.06 available at: http://manual.jointcommission.org/releases/ TJC2016A ~~TJC2015B2~~/

S.5. **Time Period for Data** *(What is the time period in which data will be aggregated for the measure, e.g., 12 mo, 3 years, look back to August for flu vaccination? Note if there are different time periods for the numerator and denominator.)*
Episode of care

S.6. **Numerator Details** *(All information required to identify and calculate the cases from the target population with the target process, condition, event, or outcome such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
IF an OUTCOME MEASURE, describe how the observed outcome is identified/counted. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
Two data elements are used for the observed outcome and to calculate the numerator:

1. ICD-10-PCS Other Procedure Codes - The International Classification of Diseases, Tenth Revision, Procedure Coding System code that identifies significant procedures performed other than the principal procedure during this hospitalization.
2. ICD-10-PCS Principal Procedure Code - The International Classification of Diseases, Tenth Revision, Procedure Coding System (ICD-10-PCS) code that identifies the principal procedure performed during this hospitalization. The principal procedure is the procedure performed for definitive treatment rather than diagnostic or exploratory purposes, or which is necessary to take care of a complication.

S.7. **Denominator Statement** *(Brief, narrative description of the target population being measured)*
The outcome target population being measured is: Nulliparous patients delivered of a live term singleton newborn in vertex presentation ICD-10-PCS Principal or Other Diagnosis Codes for delivery as defined in Appendix A, Tables 11.01.1 available at: http://manual.jointcommission.org/releases/ TJC2016A ~~TJC2015B2~~/

S.8. **Target Population Category** *(Check all the populations for which the measure is specified and tested if any):*
Maternal Health

S.9. **Denominator Details** *(All information required to identify and calculate the target population/denominator such as definitions, specific data collection items/responses , code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
~~Eight~~ Seven data elements are used to identify the outcome target population and to calculate the denominator:

1. Admission Date – The month, day and year of admission to acute inpatient care.
2. Birthdate - The month, day and year the patient was born.
~~3. Clinical Trial - Documentation that during this hospital stay the patient was enrolled in a clinical trial in which patients with pregnancy were being studied. Allowable values: Yes or No/UTD~~
3~~4~~. Discharge Date – The month day and year the patient was discharged from acute care, left against medical advice or expired during the stay.
4~~5~~. Gestational Age – Documentation of the weeks of gestation completed at the time of delivery. Allowable Values: 1-50 or UTD.
5~~6~~. ICD-10-CM Other Diagnosis Codes - The International Classification of Diseases, Tenth Revision, Clinical Modification codes associated with the secondary diagnoses for this hospitalization.
6~~7~~. ICD-9-CM Principal Diagnosis Code - The International Classification of Diseases, Tenth Revision, Clinical Modification code associated with the diagnosis established after study to be chiefly responsible for occasioning the admission of the patient for this hospitalization.
7~~8~~. Number of Previous Live Births - The number of live deliveries the patient experienced prior to current hospitalization. Allowable Values: 0-50 or UTD.
Updates available at: http://manual.jointcommission.org/releases/ TJC2016A ~~TJC2015B2~~/

S.10. **Denominator Exclusions** *(Brief narrative description of exclusions from the target population)*
• ICD-10-CM Principal Diagnosis Code or ICD-10-CM Other Diagnosis Codes for multiple gestations and other presentations as defined in Appendix A, Table 11.09
• Less than 8 years of age
• Greater than or equal to 65 years of age
• Length of Stay >120 days

• ~~Enrolled in clinical trials~~
• Gestational Age < 37 weeks or UTD

S.11. **Denominator Exclusion Details** *(All information required to identify and calculate exclusions from the denominator such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
• Patients with ICD-10-CM Principal Diagnosis Code or Other Diagnosis Codes for multiple gestations and other presentations are excluded.
• The patient age in years is equal to the Admission Date minus the Birthdate. Patients less than 8 years of age or greater or equal to 65 years of age are excluded.
• Length of stay (LOS) in days is equal to the Discharge Date minus the Admission Date. If the LOS is greater than 120 days, the patient is excluded.
~~• Patients are excluded if "Yes" is selected for Clinical Trial.~~
• Patients with a Gestational Age less than 37 weeks or UTD are excluded from the measure.

S.12. **Stratification Details/Variables** *(All information required to stratify the measure results including the stratification variables, definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format with at S.2b)*
~~The Stratification Table used for direct standardization includes the Set Number, Stratified By, and the Age Stratum (Allowable Value). The Age Stratum refers to Patient Age which is calculated by the data element Admission Date minus the data element Birthdate. Each case will be stratified according to the patient age, after the Category Assignments (e.g., numerator, denominator, not in measure population) are completed and the overall rate is calculated.~~

~~Stratified By Age Stratum~~
~~• Overall Rate No allowable value exists for the overall rate. It includes all patients greater than or equal to 8 years and less than 65 years.~~
~~• Age 8 years through 14 years A Patient Age (Admission Date minus Birthdate) greater than or equal to 8 years and less than 15 years.~~
~~• Age 15 years through 19 years A Patient Age (Admission Date minus Birthdate) greater than or equal to 15 years and less than 20 years.~~
~~• Age 20 years through 24 years A Patient Age (Admission Date minus Birthdate) greater than or equal to 20 years and less than 25 years.~~
~~• Age 25 years through 29 years A Patient Age (Admission Date minus Birthdate) greater than or equal to 25 years and less than 30 years.~~
~~• Age 30 years through 34 years A Patient Age (Admission Date minus Birthdate) greater than or equal to 30 years and less than 35 years.~~
~~• Age 35 years through 40 years A Patient Age (Admission Date minus Birthdate) greater than or equal to 35 years and less than 40 years.~~
~~• Age 40 years through 44 years A Patient Age (Admission Date minus Birthdate) greater than or equal to 40 years and less than 45 years.~~
~~• Age 45 years through 64 years A Patient Age (Admission Date minus Birthdate) greater than or equal to 45 years and less than 65 years.~~

Not Applicable

S.13. **Risk Adjustment Type** (Select type. Provide specifications for risk stratification in S.12 and for statistical model in S.14-15)
~~Other~~
If other: ~~Direct rate standardization to the distribution of the 2006 US population of nulliparous births. See attached spreadsheet for age bands used in the direct standardization.~~

S.14. **Identify the statistical risk model method and variables** *(Name the statistical method - e.g., logistic regression and list all the risk factor variables. Note - risk model development and testing should be addressed with measure testing under Scientific Acceptability)*
Not Applicable

S.15. **Detailed risk model specifications** *(must be in attached data dictionary/code list Excel or csv file. Also indicate if available at measure-specific URL identified in S.1.)*

Note: Risk model details (including coefficients, equations, codes with descriptors, definitions), should be provided on a separate worksheet in the suggested format in the Excel or csv file with data dictionary/code lists at S.2b.
~~Available in attached Excel or csv file at S.2b~~

S.15a. **Detailed risk model specifications** (*if not provided in excel or csv file at S.2b*)
~~See attachment.~~

S.16. **Type of score:**
Rate/proportion
If other:

S.17. **Interpretation of Score** *(Classifies interpretation of score according to whether better quality is associated with a higher score, a lower score, a score falling within a defined interval, or a passing score)*
Better quality = Lower score

S.18. **Calculation Algorithm/Measure Logic** *(Describe the calculation of the measure score as an ordered sequence of steps including identifying the target population; exclusions; cases meeting the target process, condition, event, or outcome; aggregating data; risk adjustment; etc.)*
1. Start processing. Run cases that are included in the PC-Mother Initial Patient Population and pass the edits defined in the Transmission Data Processing Flow: Clinical through this measure.

2. Check ICD-10-CM Principal or Other Diagnosis Codes
a. If at least one of the ICD-10-CM Principal or Other Diagnosis Code is on Table 11.09, the case will proceed to a Measure Category Assignment of B and will not be in the measure population. Stop processing.
b. If none of the ICD-10-CM Principal or Other Diagnosis Code is on Table 11.09, continue processing and proceed to recheck ICD-10-CM Principal or Other Diagnosis Codes.

3. Recheck ICD-10-CM Principal or Other Diagnosis Codes
a. If none of the ICD-10-CM Principal or Other Diagnosis Codes is on Table 11.08, the case will proceed to a Measure Category Assignment of B and will not be in the measure population. Stop processing.
b. If at least one of the ICD-10-CM Principal or Other Diagnosis Codes is on Table 11.08, continue processing and proceed to Gestational Age ~~Clinical Trial~~.

~~4. Check Clinical Trial~~
~~a. If Clinical Trial is missing, the case will proceed to a Measure Category Assignment of X and will be rejected. Stop processing.~~
~~b. If Clinical Trial equals Yes, the case will proceed to a Measure Category Assignment of B and will not be in the measure population. Stop processing.~~
~~c. If Clinical Trial equals No, continue processing and proceed to Gestational Age.~~

4~~5~~. Check Gestational Age
a. If Gestational Age is missing, the case will proceed to a Measure Category Assignment of X and will be rejected. Stop processing.
b. If Gestational Age is less than 37 or equal to a Not Unable to Determine Value, the case will proceed to a Measure Category Assignment of B and will not be in the measure population. Stop processing.
c. If Gestational Age is greater than or equal to 37, continue processing and proceed to Parity.

5~~6~~. Check Number of Previous Live Births
a. If Number of Previous Live Births is missing, the case will proceed to a Measure Category Assignment of X and will be rejected. Stop processing.
b. If Number of Previous Live Births is greater than 0, the case will proceed to a Measure Category Assignment of B for and will not be in the measure population. Stop processing.
c. If Number of Previous Live Births equals a Non Unable to Determine Value, the case will proceed to a Measure Category Assignment of E and will be in the Numerator Population. Stop processing.
d. If Parity equals 0, continue processing and proceed to check ICD-10-PCS Principal or Other Procedure Codes.

6~~7~~. Check ICD-10-PCS Principal or Other Procedure Codes
a. If all of the ICD-10-PCS Principal or Other Procedure Codes are missing or none of the ICD-10-PCS Principal or Other Procedure Codes is on Table 11.06, the case will proceed to a Measure Category Assignment of D and will be in the measure population. Stop

processing.
b. If at least one of the ICD-10-PCS Principal or Other Procedure Code is on Table 11.06, the case will proceed to a Measure Category Assignment of E and will be in the Numerator Population. Stop processing.

S.19. **Calculation Algorithm/Measure Logic Diagram URL or Attachment** *(You also may provide a diagram of the Calculation Algorithm/Measure Logic described above at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
Available at measure-specific web page URL identified in S.1

S.20. **Sampling** *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*
IF a PRO-PM, identify whether (and how) proxy responses are allowed.
The initial patient population includes patients admitted to the hospital for inpatient acute care for deliveries. Patients are included if they have: ICD-10-PCS Principal or Other Procedure Code as defined in Appendix A, Table 11.01.1, a Patient Age (Admission Date – Birthdate) >= 8 years and < 65 and a Length of Stay (Discharge Date - Admission Date) = 120 days. The sample is taken randomly as follows for a monthly sample:
• Average monthly Initial Patient Population >= 501 results in a minimum random sample size of 101.
• Average monthly Initial Patient Population 126 – 500 results in a minimum random sample size of 20% of the population size.
• Average monthly Initial Patient Population 25 – 125 results in a minimum random sample size of 25.
• Average monthly Initial Patient Population < 25 results in no sampling; 100% Initial Patient Population required

S.21. **Survey/Patient-reported data** *(If measure is based on a survey, provide instructions for conducting the survey and guidance on minimum response rate.)*
IF a PRO-PM, specify calculation of response rates to be reported with performance measure results.
Not Applicable

S.22. **Missing data** (specify how missing data are handled, e.g., imputation, delete case.)
Required for Composites and PRO-PMs.
Any file with missing data will result in a measure category assignment of X and rejection of the file from the warehouse, unless the data are not used to process the measure. If the data are used to process the measure and have been reported by the abstractor as No/UTD, the case will result in a measure category assignment of E (i.e., failed measure, not rejected).

S.23. **Data Source** *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*
If other, please describe in S.24.
Paper Medical Records

S.24. **Data Source or Collection Instrument** *(Identify the specific data source/data collection instrument e.g. name of database, clinical registry, collection instrument, etc.)*
IF a PRO-PM, identify the specific PROM(s); and standard methods, modes, and languages of administration.
Each data element in the data dictionary includes suggested data sources. The data are collected using contracted Performance Measurement Systems (vendors) that develop data collection tools based on the measure specifications. The tools are verified and tested by Joint Commission staff to confirm the accuracy and conformance of the data collection tool with the measure specifications. The vendor may not offer the measure set to hospitals until verification has been passed.

S.25. **Data Source or Collection Instrument** *(available at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
No data collection instrument provided

S.26. **Level of Analysis** *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*
Facility, Population : National
S.27. **Care Setting** *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*
Hospital/Acute Care Facility
If other:

S.28. **COMPOSITE Performance Measure** - Additional Specifications *(Use this section as needed for aggregation and weighting rules, or calculation of individual performance measures if not individually endorsed.)*
Not Applicable

2a. Reliability – See attached Measure Testing Submission Form

2b. Validity – See attached Measure Testing Submission Form
PC-02_0471_MeasureTesting_MSF5.0_Data.doc

NATIONAL QUALITY FORUM

Measure missing data in MSF 6.5 from MSF 5.0

NQF #: 0471 **NQF Project:** Perinatal and Reproductive Health Project

2. RELIABILITY & VALIDITY - SCIENTIFIC ACCEPTABILITY OF MEASURE PROPERTIES

Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. **(evaluation criteria)**
Measure testing must demonstrate adequate reliability and validity in order to be recommended for endorsement. Testing may be conducted for data elements and/or the computed measure score. Testing information and results should be entered in the appropriate field. Supplemental materials may be referenced or attached in item 2.1. See guidance on measure testing.

2a2. Reliability Testing. (*Reliability testing was conducted with appropriate method, scope, and adequate demonstration of reliability.*)

2a2.1 Data/Sample *(Description of the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)*:
The PC measure set has been in national use since the 2nd quarter of 2010. It is a requirement of participation in the ORYX initiative that data on all measures in the set are collected. (ORYX is the term used by The Joint Commission to describe the component of the hospital accreditation program which requires data collection and reporting on standardized national performance measures.)
Demographics of organizations collecting and reporting data on these measures are as follows:
163 health care organizations representing various types, locations and sizes:
10 For Profit, 91 Not for Profit, 46 Military Facilities, 9 County, 2 State, 5 Other
15 >=500 beds; 29 250-499 beds; 50 100-249 beds; 69 <100 beds
Located in: AE, AK, AL, AP, AR, AZ, CA, DO, DC, FL, GA, HI, ID, IL, IN, KS, KY, LA, MA, MD, MI, MN, MO, MS, MT, NC, NE, NV, NY, OH,OK, PA, PR, RI, SC, TN, TX, VA, WA, WI, WV
26 performance measurement systems

2a2.2 Analytic Method *(Describe method of reliability testing & rationale)*:
This measure was adapted from NQF-endorsed measure 0471 Cesarean Rate for Low-Risk First Birth Women (NTSV CS Rate). As such, initial data reliability would have been addressed during the original endorsement. The Joint Commission will be conducting additional reliability studies on this measure as well as the entire PC measure set beginning in October 2011.

.
Currently, hospitals are supported in their data collection and reporting efforts by 26 contracted performance measurement system (PMS) vendors. It is a contractual requirement of Joint Commission listed vendors that the quality and reliability of data submitted to them by contracted health care organizations must be monitored on a quarterly basis. In addition, The Joint Commission analyzes these data by running 17 quality tests on the data submitted into ORYX. (ORYX is the term used by The Joint Commission to describe the component of the hospital accreditation program which requires data collection and reporting on standardized national performance measures). The following is a list of the major tests done on the submitted ORYX data, taken from the 2011 ORYX Performance Measurement System Requirements manual.

- Transmission of complete data
- Usage of individual core measure data received: To understand if the HCO provides the relevant service to treat the relevant population
- Investigation of aberrant data points

- Verification of patient population and sample size
- Identification of missing data elements
- Validation of the accuracy of target outliers
- Data integrity
- Data corrections

Data Element Agreement Rate:
Inter-rater reliability testing methodology utilized by contracted performance measure system vendors as outlined in the contract is as follows:

- All clinical data elements and all editable demographic elements are scored.
- All measure data are reabstracted with originally abstracted data having been blinded so that the reabstraction is not biased.
- Reabstracted data are compared with originally abstracted data on a data element by data element basis. A data element agreement rate is calculated. Clinical and demographic data are scored separately, and an overall agreement rate is computed.

2a2.3 Testing Results *(Reliability statistics, assessment of adequacy in the context of norms for the test conducted)***:**
Data element agreement rates were reported to The Joint Commission for 1Q11. This reflects the findings of 108 hospitals, comprising 13,279 records (100% sample). The following table delineates calculated agreement rates for individual data elements that are used to compute measure rates for PC-02.

Data Elements with a Mismatch - Mother	total n	total d	rate
Gestational Age	639	712	89.75%
Parity	492	505	97.43%

These agreement rates are considered to be well within acceptable levels.

2b. VALIDITY. Validity, Testing, <u>including all Threats to Validity</u>: H☐ M☐ L☐ I ☐

2b1.1 Describe how the measure specifications *(measure focus, target population, and exclusions)* **are consistent with the evidence cited in support of the measure focus (***criterion 1c)* **and identify any differences from the evidence:**
This measure focuses on reducing the number of cesarean sections performed on nulliparous patients delivering a live term singleton newborn in vertex presentation. The literature supports the focus on nulliparous patients with live term singleton newborns in vertex position delivering vaginally. Accordingly, this measure focuses on patients without common conditions that have accepted high rates of cesarean birth, i.e., breech presentation, multiple gestations. Also excluded from the measure are patients with a length of stay greater than 120 days, and those enrolled in a clinical trial. These exclusions are not addressed in the literature, but are included for this measure in order to harmonize with other CMS/Joint Commission aligned measures.

2b2. Validity Testing. (*Validity testing was conducted with appropriate method, scope, and adequate demonstration of validity.)*

2b2.1 Data/Sample *(Description of the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)***:**
As previously mentioned the PC measure set has been in national use since the 2nd quarter of 2010. Demographics of organizations collecting and reporting data on these measures are as follows:
163 health care organizations representing various types, locations and sizes:
10 For Profit, 91 Not for Profit, 46 Military Facilities, 9 County, 2 State, 5 Other
15 >=500 beds; 29 250-499 beds; 50 100-249 beds; 69 <100 beds
Located in: AE, AK, AL, AP, AR, AZ, CA, DO, DC, FL, GA, HI, ID, IL, IN, KS, KY, LA, MA, MD, MI, MN, MO, MS, MT, NC, NE, NV, NY, OH,OK, PA, PR, RI, SC, TN, TX, VA, WA, WI, WV
26 performance measurement systems

2b2.2 Analytic Method *(Describe method of validity testing and rationale; if face validity, describe systematic assessment)***:**
Since the measure has been in national use, continued face validity of the measure has been determined through analysis of feedback from measure users. The Joint Commission provides a web-based application with which measure users can provide feedback regarding appropriateness of measure specifications, request clarification of specifications, and/or provide other comments pertinent to the measure. This feedback is systematically continually reviewed in order to identify trends and to identify areas of the measure specifications that require clarification or revision. Additionally, Joint Commission staff continually monitors the national literature and environment in order to assess continued validity of this measure.

As noted previously, The Joint Commission is currently performing reliability site visits this year. A component of these visits will include focus group interviews with hospital staff working with the PC measures to obtain feedback regarding the validity of the

measures and suggestions for further refinement of the specifications.

2b2.3 Testing Results *(Statistical results, assessment of adequacy in the context of norms for the test conducted; if face validity, describe results of systematic assessment)*:
Analysis of feedback obtained via our automated feedback system reveals slightly more than 80 submissions regarding specifications for this measure since its implementation in 2010. Predominant themes of these submissions involved questions regarding clarification of the data elements Parity and Gestational Age with respect to both definitions and the calculation of gestational age, the order of priority sources to retrieve the data and incorporation of GTPAL terminology for Parity. Additional notes for abstractors were added to the data elements for clarification. An additional ICD-9-CM diagnosis code identifying footling breech was also added to Table 11.09 to update exclusions based on consultation with the original measure developer.

POTENTIAL THREATS TO VALIDITY. (*All potential threats to validity were appropriately tested with adequate results.*)

2b3. Measure Exclusions. (*Exclusions were supported by the clinical evidence in 1c or appropriately tested with results demonstrating the need to specify them.*)

2b3.1 Data/Sample for analysis of exclusions *(Description of the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)*:
As previously mentioned the PC measure set has been in national use since the 2nd quarter of 2010. Demographics of organizations collecting and reporting data on these measures are as follows:
163 health care organizations representing various types, locations and sizes:
10 For Profit, 91 Not for Profit, 46 Military Facilities, 9 County, 2 State, 5 Other
15 >=500 beds; 29 250-499 beds; 50 100-249 beds; 69 <100 beds

Located in: AE, AK, AL, AP, AR, AZ, CA, DO, DC, FL, GA, HI, ID, IL, IN, KS, KY, LA, MA, MD, MI, MN, MO, MS, MT, NC, NE, NV, NY, OH,OK, PA, PR, RI, SC, TN, TX, VA, WA, WI, WV
26 performance measurement systems

2b3.2 Analytic Method (*Describe type of analysis and rationale for examining exclusions, including exclusion related to patient preference)*:
Measure exclusions that were not derived directly from the evidence are presented below. Please note that these are population exclusions that are necessary to ensure consistency in all measures in this 5 measure set.
These exclusions were analyzed for frequency of occurrence. An issue that is of great concern to users of this measure is that due to the presence of exceptions to the measure, attainment of a 0% measure rate is not possible. Because of the role of this measure in the current Joint Commission accreditation process, this is especially troubling to measure users. This concern is the basis for the non-evidence-based exclusions to these measures. Additional reasons for these population exclusions are enumerated in our response to section 2b1.1 above. The measure exclusions that were not derived directly from the evidence are as follows:

1. Patients with LOS <120 days
2. Patients less than 8 years of age or greater than or equal to 65 years of age
3. Patients enrolled in clinical trials

2b3.3 Results *(Provide statistical results for analysis of exclusions, e.g., frequency, variability, sensitivity analyses):*
N=353,671
1. Patients who have a length of stay (LOS) greater than 120 days =0%
2. Patients less than 8 years of age or greater than or equal to 65 years of age=0%
3. Patients enrolled in clinical trials =0.04%

2b4. Risk Adjustment Strategy. (*For outcome measures, adjustment for differences in case mix (severity) across measured entities was appropriately tested with adequate results.*)

2b4.1 Data/Sample *(Description of the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)*:
Not Applicable

2b4.2 Analytic Method (*Describe methods and rationale for development and testing of risk model or risk stratification including selection of factors/variables)*:
Not Applicable

2b4.3 Testing Results *(Statistical risk model: Provide quantitative assessment of relative contribution of model risk factors; risk model performance metrics including cross-validation discrimination and calibration statistics, calibration curve and risk decile plot, and assessment of adequacy in the context of norms for risk models. Risk stratification: Provide quantitative assessment of relationship of risk factors to the outcome and differences in outcomes among the strata)*:
Not Applicable

2b4.4 If outcome or resource use measure is not risk adjusted, provide rationale and analyses to justify lack of adjustment:
Not Applicable

2b5. Identification of Meaningful Differences in Performance. (*The performance measure scores were appropriately analyzed and discriminated meaningful differences in quality.*)

2b5.1 Data/Sample *(Describe the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)*:
As previously mentioned the PC measure set has been in national use since the 2nd quarter of 2010. Demographics of organizations collecting and reporting data on these measures are as follows:
163 health care organizations representing various types, locations and sizes:
10 For Profit, 91 Not for Profit, 46 Military Facilities, 9 County, 2 State, 5 Other
15 >=500 beds; 29 250-499 beds; 50 100-249 beds; 69 <100 beds
Located in: AE, AK, AL, AP, AR, AZ, CA, DO, DC, FL, GA, HI, ID, IL, IN, KS, KY, LA, MA, MD, MI, MN, MO, MS, MT, NC, NE, NV, NY, OH,OK, PA, PR, RI, SC, TN, TX, VA, WA, WI, WV
26 performance measurement systems

2b5.2 Analytic Method *(Describe methods and rationale to identify statistically significant and practically/meaningfully differences in performance)*:
The method used to analyze meaningful differences in performance at The Joint Commission is Target Analysis. The object of target analysis is to compare a health care organization's (HCO) data against a comparative norm for the purpose of evaluating performance improvement opportunities. When an organization's performance level is statistically significantly different from a comparative norm, it is considered a statistical deviation. A statistical deviation may be desirable or undesirable depending on the "direction of improvement" of the measure.
There are two components to the target analysis methodology used at The Joint Commission. Given the national average for a performance measure, a target range is constructed. Using generalized linear mixed models methodology (also known as hierarchical models), a predicted estimate of an HCO's performance, with a corresponding 95% confidence interval, is generated. This confidence interval is compared to the target range, to determine the HCOs' rating. The estimate of the organization's true performance is based on both the data from that organization and on data from the entire set of reporting organizations.

2b5.3 Results *(Provide measure performance results/scores, e.g., distribution by quartile, mean, median, SD, etc.; identification of statistically significant and meaningfully differences in performance)*:
PC-02 Distribution of Outliers

2011 1st Quarter Data:
Scores on this measure: N=160, Mean 26.7%, SD 0.12953
10th Percentile= 14%
25th Percentile= 19.4%
50th Percentile=26%
75th Percentile= 32.5%
90th Percentile= 40%

159 (100%) Neutral – results not significantly different from target range

2b6. Comparability of Multiple Data Sources/Methods. (*If specified for more than one data source, the various approaches result in comparable scores.*)

2b6.1 Data/Sample *(Describe the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)*:
Multiple data sources are not used for this measure.

2b6.2 Analytic Method *(Describe methods and rationale for testing comparability of scores produced by the different data sources specified in the measure)*:
Not Appliocable

2b6.3 Testing Results *(Provide statistical results, e.g., correlation statistics, comparison of rankings; assessment of adequacy in the context of norms for the test conducted)*:
Not Appliocable

2c. Disparities in Care: H☐ M☐ L☐ I ☐ NA☐ (*If applicable, the measure specifications allow identification of disparities.*)

2c.1 If measure is stratified for disparities, provide stratified results *(Scores by stratified categories/cohorts)*: The measure is not stratified for disparities.

2c.2 If disparities have been reported/identified (e.g., in 1b), but measure is not specified to detect disparities, please explain:
Although a rise in cesarean sections is present for all racial groups with the highest increase in African Americans, there are no plans to stratify this measure.. The Joint Commission does not currently capture date elements for race or ethnicity because these data elements have not been shown to be reliably collectable due to the fact that no national standardized definitions exist for these data elements. Also, not all hospitals collect race and ethnicity. In the future, it may be feasible for The Joint Commission to explore how race and ethnicity and other relevant disparity data, might be collected reliably in the future. The measure is currently stratified by age groups capturing advanced maternal age. The data from the different age groups are used in the direct standardization model applied to each hospital's rate.

2.1-2.3 Supplemental Testing Methodology Information:

Steering Committee: Overall, was the criterion, *Scientific Acceptability of Measure Properties*, met?
(*Reliability and Validity must be rated moderate or high*) Yes☐ No☐
Provide rationale based on specific subcriteria:

If the Committee votes No, STOP

NEW (since 2012 endorsement): Data for Empirical Testing

Validity Testing Section PC-02
2b2. VALIDITY TESTING
2b2.1. What level of validity testing was conducted? (*may be one or both levels*)

- ☐ **Critical data elements** (*data element validity must address ALL critical data elements*)
- ☐ **Performance measure score**
- ☒ **Empirical validity testing**
- ☐ **Systematic assessment of face validity of performance measure score as an indicator** of quality or resource use (*i.e., is an accurate reflection of performance on quality or resource use and can distinguish good from poor performance*)

2b2.2. For each level of testing checked above, describe the method of validity testing and what it tests (*describe the steps—do not just name a method; what was tested, e.g., accuracy of data elements compared to authoritative source, relationship to another measure as expected; what statistical analysis was used)*

ICD-9 to ICD-10 Conversion Process:
The goal was to convert ICD-9 to ICD-10 equivalent codes, consistent with the clinical intent of the original measure specifications. The Joint Commission worked with a certified coding expert throughout the conversion process. The 3M Coding Conversion Tool was utilized, including forward mapping of ICD-9 codes to ICD-10 codes as well as reverse mapping from ICD-10 to ICD-9 to ensure appropriateness. MSDRGs and instructions in the tabular index were also examined to ensure appropriate code mapping. Crosswalks comprising ICD-9

codes mapped to ICD-10 codes were created and reviewed by members of the Technical Advisory Panel, CMS subcontractors, and performance measurement system vendors prior to being posted for a 12 month public comment period. Feedback from the field indicated that the crosswalks generally were mapped correctly. Minor modifications to the code tables were made as needed. Final code tables were published in early 2015, well in advance of the mandated date of October 1, 2015.
Perinatal Care (PC) Initial Patient Population
The PC measure set is unique in that there are two distinct Initial Patient Populations within the measure set, mothers (PC-01, PC-02, PC-03) and newborns. (PC-04, PC-05).

Subpopulation Mothers
Patients admitted to the hospital for inpatient acute care are included in the PC Mother Initial sampling group if they have: ICD-9-CM Principal or Other Diagnosis Code as defined in Appendix A, Tables 11.01, 11.02, 11.03, or 11.04, a Patient Age (Admission Date Birthdate) >= 8 years and < 65 and a Length of Stay (Discharge Date - Admission Date) ≤ 120 days
PC-02 - Cesarean Section belongs to the above population.

The data used to measure the validity of the PC measure are comprised of data from the third and fourth quarters of 2014, and the first and second quarters of 2015. 1,345 hospitals submitted 2,695,467 inpatient records for all the elected PC measures. The hospitals included in the analysis reported one year of data and had 30 or more denominator cases in the analysis period.
Measure convergent validity for PC-02 was assessed using hospitals patient level data from The Joint commission warehouse. Measure specifications, including population identification, numerator and denominator statements, exclusions, and data elements and their definitions were found to be understandable, retrievable, and relevant in previous validity testing.

2b2.3. What were the statistical results from validity testing? (*e.g., correlation; t-test*)
Overall descriptive statistics for sub population MOTHER:

Note: The observed rates were used for PC-02 in the analysis in this section; the PC-02 hospital rates were not age standardized.

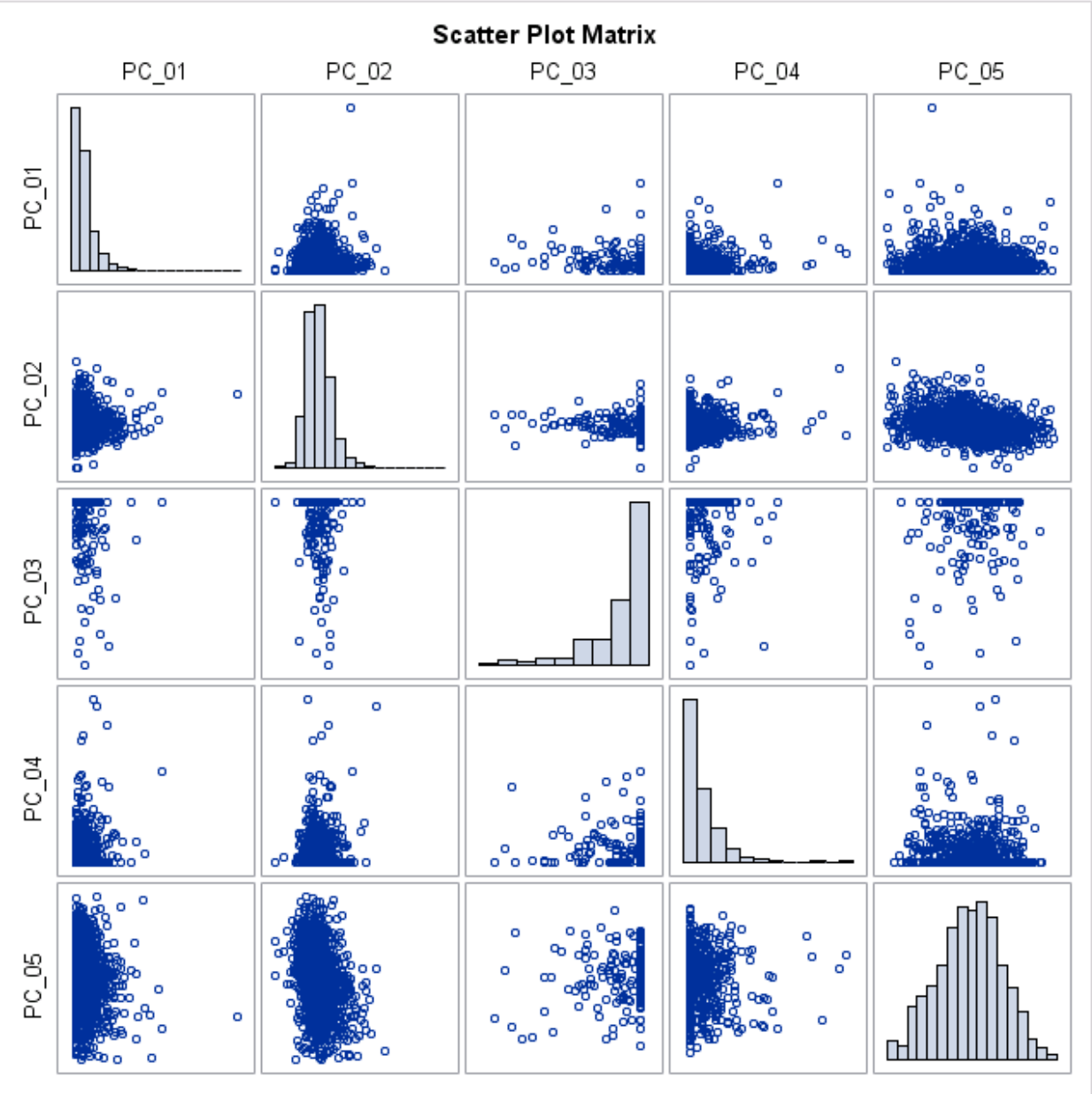
N=1,345 hospitals
n = 2,695,467 records submitted

Descriptive statistics for PC-02 measure:
N=1,345 hospitals
n = 1,169,924

Min = 0%
Mean: 26.2%
Percentile 10%: 18%
Percentile 25%: 21%
Median: 25.4%
Percentile 75%: 30.4%
Percentile 90%: 36%
Max = 100%

Simple Statistics						
Variable	N	Mean	Std Dev	Median	Minimum	Maximum
PC_01	1237	0.02753	0.03803	0.01734	0	0.51240
PC_02	1345	0.26287	0.07974	0.25410	0	1.00000
PC_03	162	0.97762	0.03311	0.99425	0.84615	1.00000
PC_04	523	0.05267	0.08432	0.02203	0	0.66129
PC_05	1352	0.49198	0.19284	0.50190	0.00317	1.00000

Spearman Correlation Coefficients Prob > |r| under H0: Rho=0 Number of Observations					
	PC_01	PC_02	PC_03	PC_04	PC_05
PC_01	1.00000 1237	0.06843 0.0163 1231	-0.26960 0.0006 159	0.10724 0.0169 496	-0.03538 0.2137 1237
PC_02	0.06843 0.0163 1231	1.00000 1345	-0.18318 0.0196 162	0.02807 0.5218 523	-0.32009 <.0001 1343
PC_03	-0.26960 0.0006 159	-0.18318 0.0196 162	1.00000 162	-0.03117 0.7030 152	0.07729 0.3283 162
PC_04	0.10724 0.0169 496	0.02807 0.5218 523	-0.03117 0.7030 152	1.00000 523	-0.03560 0.4165 523
PC_05	-0.03538 0.2137 1237	-0.32009 <.0001 1343	0.07729 0.3283 162	-0.03560 0.4165 523	1.00000 1352



The Spearman rank-order correlation is a nonparametric measure of association based on the ranks of the data values by measure PC-02 and hospitals. We used this methodology because of the skewness of the distribution of the measure rates.

2b2.4. What is your interpretation of the results in terms of demonstrating validity? (i.*e., what do the results mean and what are the norms for the test conducted?*)

The correlation of PC-02 with the other PC measures in the PC measure set indicates that the correlations (with the exception of PC-04), although in the expected direction and statistically significant, are relatively weak. Although 90% of the hospital measure rates fall between 18 and 36%, there are still a number of hospitals with measure rates significantly greater than 36% and less than 18%, indicating that the performance of hospitals on this measure are not uniformly acceptable.

2b3. EXCLUSIONS ANALYSIS .
NA ☐ no exclusions — *skip to section 2b4*

2b3.1. Describe the method of testing exclusions and what it tests (*describe the steps—do not just name a method; what was tested, e.g., whether exclusions affect overall performance scores; what statistical analysis was used*)

There were 1,169,924 admissions selected from the initial cohort. From among the 1,169,924 admissions in 1,345 hospitals, the descriptive statistics are given below.

The following exclusions were analyzed by subpopulation and measure for frequency and variability across providers:

- *ICD-9-CM Principal Diagnosis Code or ICD-9-CM Other Diagnosis Codes* for multiple gestations and other presentations as defined in Appendix A, Table 11.09
- Less than 8 years of age
- Greater than or equal to 65 years of age
- Length of Stay >120 days
- Enrolled in clinical trials
- *Gestational Age* < 37 weeks or UTD

2b3.2. What were the statistical results from testing exclusions? (*include overall number and percentage of individuals excluded, frequency distribution of exclusions across measured entities, and impact on performance measure scores*)

Exclusion Subpopulation 1 – PC-02:
ICD-9-CM Principal Diagnosis Code or ICD-9-CM Other Diagnosis Codes for multiple gestations and other presentations as defined in Appendix A, Table 11.09:
Exclusion: No observations noted

Less than 8 years of age
Exclusion: Included in the initial population exclusion

Greater than or equal to 65 years of age
Exclusion*:* Included in the initial population exclusion

Length of Stay >120 days
Exclusion: Included in the initial population exclusion

Exclusion: Patients enrolled in clinical trials
Overall Occurrence n =729
Overall Occurrence Percentage: 0.06%
Minimum: 0%

10th Percentile: 0%
Median: 0%
90th Percentile: 0.03%
Maximum: 7.97%

Exclusion: Gestational Age < 37 weeks or UTD
Overall Number of Occurrences n = 113,520
Overall Occurrence Percentage: 9.7%
Minimum: 0.29%
10th Percentile: 5%
Median: 8.7%
90th Percentile: 14.8%
Maximum: 34%

2b3.3. What is your interpretation of the results in terms of demonstrating that exclusions are needed to prevent unfair distortion of performance results? (*i.e., the value outweighs the burden of increased data collection and analysis. Note: **If patient preference is an exclusion**, the measure must be specified so that the effect on the performance score is transparent, e.g., scores with and without exclusion*)

The difference between the 10th and 90th percentiles of the distribution of exclusion rates is narrow indicating that the occurrence is random and likely would not bias performance results.

It is believed that all of the exclusions should be retained for the following reasons:

Exclusion: *ICD-9-CM Principal Diagnosis Code or ICD-9-CM Other Diagnosis Codes* for multiple gestations and other presentations as defined in Appendix A, Table 11.09:
Rationale: *Table 11.09 contains diagnosis codes for a fetus in any position other than a vertex position and any gestations of two more in order to exclude these cases from the denominator.*

Exclusion: Patients who have a *Length of stay* greater than 120 days
Rationale: Included for this measure in order to harmonize with other CMS/Joint Commission aligned measures.

Exclusion: Patients with Gestational Age < 37 weeks or UTD
Rationale: The denominator population is limited to patients ≥ 37 or more weeks of completed gestation. Patients with UTD for gestational age typically have had no prenatal care.

Note:
After the submission of this form and before the standing committee meeting at which this measure was discussed the Joint Commission's Perinatal Care Technical Advisory Panel recommended Clinical Trial be removed as a denominator exclusion since this is a rare occurrence.

2b4. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES
If not an intermediate or health outcome, or PRO-PM, or resource use measure, skip to section 2b5.

2b4.1. What method of controlling for differences in case mix is used?

☒ **No risk adjustment or stratification**
☐ **Statistical risk model with** Click here to enter number of factors **risk factors**
☐ **Stratification by** Click here to enter number of categories **risk categories**
☐ **Other,**

2b4.2. If an outcome or resource use measure is not risk adjusted or stratified, provide rationale and analyses to demonstrate that controlling for differences in patient characteristics (case mix) is not needed to achieve fair comparisons across measured entities.

Note:
After the submission of this form and before the standing committee meeting at which this measure was discussed the Joint Commission's Perinatal Care Technical Advisory Panel recommended using the simple cesarean birth rates without further risk adjustment. The decision to remove all risk-adjustment from this measure was made based on analysis of data on this measure received by The Joint Commission which indicates that age is only a weak predictor of outcome and that age standardization could potentially distort the age-standardized measure rates for hospitals with small sample sizes. Additionally, the Technical Advisory Panel considered evidence from two recent studies [1], [2] when making the recommendation to remove age standardization from the measure. Therefore, effective with discharges beginning July 1, 2016, The Joint Commission has removed all risk adjustments.

[1] Caceres IA, Arcaya M, Declercq E, Belanoff CM, Janakiraman V, et al. (2013) Hospital Differences in Cesarean Deliveries in Massachusetts (US) 2004–2006:The Case against Case-Mix Artifact. PLoS ONE 8(3): e57817. doi:10.1371/journal.pone.0057817

[2] Main E. (2014) Nuliparous, Term, Singleton, Vertex (NTSV) Cesarean Birth Rates: extreme hospital variation is not changed by adjustment for case-mix. Oral Presentation: Pacific Coast Obstetrics and Gynecology Society

2b4.3. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors (clinical factors or sociodemographic factors) used in the statistical risk model or for stratification by risk (*e.g., potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of p<0.10; correlation of x or higher; patient factors should be present at the start of care*)

2b4.4a. What were the statistical results of the analyses used to select risk factors?

2b4.4b. Describe the analyses and interpretation resulting in the decision to select SDS factors (e.g. prevalence of the factor across measured entities, empirical association with the outcome, contribution of unique variation in the outcome, assessment of between-unit effects and within-unit effects)

2b4.5. Describe the method of testing/analysis used to develop and validate the adequacy of the statistical model or stratification approach (*describe the steps—do not just name a method; what statistical analysis was used*)
Provide the statistical results from testing the approach to controlling for differences in patient characteristics (case mix) below.
If stratified, skip to 2b4.9
Not Applicable

2b4.6. Statistical Risk Model Discrimination Statistics (*e.g., c-statistic, R-squared*)**:**
Not Applicable

2b4.7. Statistical Risk Model Calibration Statistics (*e.g., Hosmer-Lemeshow statistic*):
Not Applicable

2b4.8. Statistical Risk Model Calibration – Risk decile plots or calibration curves:
Not Applicable

2b4.9. Results of Risk Stratification Analysis:

2b4.10. What is your interpretation of the results in terms of demonstrating adequacy of controlling for differences in patient characteristics (case mix)? (i.e., *what do the results mean and what are the norms for the test conducted*)

3. Feasibility

Extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

3a. Byproduct of Care Processes

For clinical measures, the required data elements are routinely generated and used during care delivery (e.g., blood pressure, lab test, diagnosis, medication order).

3a.1. **Data Elements Generated as Byproduct of Care Processes.**

generated by and used by healthcare personnel during the provision of care, e.g., blood pressure, lab value, medical condition, Coded by someone other than person obtaining original information (e.g., DRG, ICD-9 codes on claims), Abstracted from a record by someone other than person obtaining original information (e.g., chart abstraction for quality measure or registry)

If other:

3b. Electronic Sources

The required data elements are available in electronic health records or other electronic sources. If the required data are not in electronic health records or existing electronic sources, a credible, near-term path to electronic collection is specified.

3b.1. **To what extent are the specified data elements available electronically in defined fields?** *(i.e., data elements that are needed to compute the performance measure score are in defined, computer-readable fields)*

Some data elements are in defined fields in electronic sources

3b.2. **If ALL the data elements needed to compute the performance measure score are not from electronic sources, specify a credible, near-term path to electronic capture, OR provide a rationale for using other than electronic sources.**

The Joint Commission recognizes that not all hospitals currently have the capacity to abstract the electronic version of this measure, so continues to offer this chart abstracted version which allows for data capture from unstructured data fields. All data elements needed to compute the PC-02 performance measure score are in the process of being retooled for capture from electronic sources with testing scheduled to take place in 2016. Annual updates will be performed to match the eCQM specifications to the current version of the chart-abstracted specifications.

3b.3. **If this is an eMeasure, provide a summary of the feasibility assessment in an attached file or make available at a measure-specific URL.**

Attachment:

3c. Data Collection Strategy

Demonstration that the data collection strategy (e.g., source, timing, frequency, sampling, patient confidentiality, costs associated with fees/licensing of proprietary measures) can be implemented (e.g., already in operational use, or testing demonstrates that it is ready to put into operational use). For eMeasures, a feasibility assessment addresses the data elements and measure logic and demonstrates the eMeasure can be implemented or feasibility concerns can be adequately addressed.

3c.1. **Describe what you have learned/modified as a result of testing and/or operational use of the measure regarding data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, time and cost of data collection, other feasibility/implementation issues.**

IF a PRO-PM, consider implications for both individuals providing PROM data (patients, service recipients, respondents) and those whose performance is being measured.

At the present time, hospitals using this performance measure generally collect measure data via manual review of the paper medical record, the EMR or a combination of both. Collected data are submitted to The Joint Commission on a quarterly basis, by way of contracted performance measurement system vendors, as described previously. Specifications for this measure are freely available to anyone who wishes to use the measure. Feedback from hospitals using this measure indicates that required data elements are generally available in the medical record, and measure specifications are robust and easy to understand. As described above, as feedback from measure users has indicated the need for clarification or revision of measure specifications, this has taken place. The Joint Commission added Vital Records as an additional data source in the current measure specifications.

3c.2. **Describe any fees, licensing, or other requirements to use any aspect of the measure as specified** *(e.g., value/code set, risk model, programming code, algorithm)*.

There are no fees or licensing requirements to use the Joint Commission performance measures, all of which are in the public domain.

4. Usability and Use

Extent to which potential audiences (e.g., consumers, purchasers, providers, policy makers) are using or could use performance results for both accountability and performance improvement to achieve the goal of high-quality, efficient healthcare for individuals or populations.

4a. Accountability and Transparency

Performance results are used in at least one accountability application within three years after initial endorsement and are publicly reported within six years after initial endorsement (or the data on performance results are available). If not in use at the time of initial endorsement, then a credible plan for implementation within the specified timeframes is provided.

4.1. **Current and Planned Use**

NQF-endorsed measures are expected to be used in at least one accountability application within 3 years and publicly reported within 6 years of initial endorsement in addition to performance improvement.

Planned	Current Use (for current use provide URL)
Public Reporting	Regulatory and Accreditation Programs Hospital Accreditation Program http://jointcommission.org Quality Improvement with Benchmarking (external benchmarking to multiple organizations) America's Hospitals: Improving Quality and Safety – The Joint Commission's Annual Report http://www.jointcommission.org/annualreport.aspx Quality Improvement (Internal to the specific organization) Perinatal Care Certification http://www.jointcommission.org/certification/perinatal_care_certification.aspx

4a.1. **For each CURRENT use, checked above, provide:**

- Name of program and sponsor
- Purpose
- Geographic area and number and percentage of accountable entities and patients included

- Name of program and sponsor Hospital Accreditation Program; The Joint Commission
- Purpose: An accreditation program that recognizes hospitals that meet standard requirements to provide safe and effective patient care.
- Geographic area and number and percentage of accountable entities and patients included Nationwide; 3300 Joint Commission-accredited hospitals (2014)

- Name of program and sponsor America's Hospitals: Improving Quality and Safety – The Joint Commission's Annual Report ; The Joint Commission
- Purpose: The annual report summarizes the performance of Joint Commission-accredited hospitals on 46 accountability measures of evidence-based care processes closely linked to positive patient outcomes, and provides benchmarks from Top Performer on Key Quality Measures® hospitals.
- Geographic area and number and percentage of accountable entities and patients included Nationwide; 3300 Joint Commission-accredited hospitals (2014)

- Name of program and sponsor Perinatal Care Certification; The Joint Commission
- Purpose: A certification program that recognizes hospitals that have achieved integrated, coordinated, patient-centered care for clinically uncomplicated pregnancies and births.
- Geographic area and number and percentage of accountable entities and patients included Nationwide; Ten Joint Commission-accredited hospitals (2015)

4a.2. **If not currently publicly reported OR used in at least one other accountability application (e.g., payment program, certification, licensing) what are the reasons?** (*e.g., Do policies or actions of the developer/steward or accountable entities restrict access to performance results or impede implementation?*)
Not Applicable

4a.3. **If not currently publicly reported OR used in at least one other accountability application, provide a credible plan for implementation within the expected timeframes -- any accountability application within 3 years and publicly reported within 6 years of initial endorsement.** (*Credible plan includes the specific program, purpose, intended audience, and timeline for implementing the measure within the specified timeframes. A plan for accountability applications addresses mechanisms for data aggregation and reporting.*)
Not Applicable

4b. Improvement

Progress toward achieving the goal of high-quality, efficient healthcare for individuals or populations is demonstrated. If not in use for performance improvement at the time of initial endorsement, then a credible rationale describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.

4b.1. **Progress on Improvement. (Not required for initial endorsement unless available.)**
Performance results on this measure (current and over time) should be provided in 1b.2 and 1b.4. Discuss:

- **Progress (trends in performance results, number and percentage of people receiving high-quality healthcare)**
- **Geographic area and number and percentage of accountable entities and patients included**

Although the rate of NTSV cesarean births has remained relatively the same with 26.7% reported from 165 hospitals in 2010 and 26.8% reported from 1388 hospitals for 2014, improvement was noted in 2014 for the lower quartile (21.1%) and 10th percentile (17.6%) hospitals. This underscores the importance to continue to monitor progress towards improving the rate in order to reach the Healthy People 2020 goal of 23.9%. There are National efforts currently in place to develop a "toolkit" in order to help hospitals improve performance on this measure.

4b.2. **If no improvement was demonstrated, what are the reasons? If not in use for performance improvement at the time of initial endorsement, provide a credible rationale that describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.**
Not Applicable

4c. Unintended Consequences

The benefits of the performance measure in facilitating progress toward achieving high-quality, efficient healthcare for individuals or populations outweigh evidence of unintended negative consequences to individuals or populations (if such evidence exists).

4c.1. **Were any unintended negative consequences to individuals or populations identified during testing; OR has evidence of unintended negative consequences to individuals or populations been reported since implementation? If so, identify the negative unintended consequences and describe how benefits outweigh them or actions taken to mitigate them.**
Unintended Consequence:
Patients who did not receive prenatal care were inappropriately included in the measure denominator, as the gestational age data element was abstracted as unable to be determined (UTD).

Mitigating Action: In order to avoid penalizing hospitals, cases with UTD were removed from the measure population.

Unintended Consequence:
Some hospitals have reported higher rates due to small denominator populations as a result of sampling.
Mitigating Action:
Vital Records reports, delivery logs and clinical information systems were added as acceptable data sources to help hospitals identify all cases with =>37 weeks gestation, so that 100% of these cases could be reviewed to increase the denominator population size.

5. Comparison to Related or Competing Measures

If a measure meets the above criteria and there are endorsed or new related measures (either the same measure focus or the same target population) or competing measures (both the same measure focus and the same target population), the measures are compared to address harmonization and/or selection of the best measure.

5. Relation to Other NQF-endorsed Measures
Are there related measures (conceptually, either same measure focus or target population) or competing measures (conceptually both the same measure focus and same target population)? If yes, list the NQF # and title of all related and/or competing measures.
No

5.1a. **List of related or competing measures (selected from NQF-endorsed measures)**

5.1b. **If related or competing measures are not NQF endorsed please indicate measure title and steward.**

5a. Harmonization

The measure specifications are harmonized with related measures;
OR
The differences in specifications are justified

5a.1. **If this measure conceptually addresses EITHER the same measure focus OR the same target population as NQF-endorsed measure(s):**
Are the measure specifications completely harmonized?

5a.2. **If the measure specifications are not completely harmonized, identify the differences, rationale, and impact on interpretability and data collection burden.**
Not Applicable

5b. Competing Measures

The measure is superior to competing measures (e.g., is a more valid or efficient way to measure);
OR
Multiple measures are justified.

5b.1. **If this measure conceptually addresses both the same measure focus and the same target population as NQF-endorsed measure(s):**
Describe why this measure is superior to competing measures (e.g., a more valid or efficient way to measure quality); OR provide a rationale for the additive value of endorsing an additional measure. (Provide analyses when possible.)
Not Applicable

Appendix

A.1 Supplemental materials may be provided in an appendix. All supplemental materials (such as data collection instrument or methodology reports) should be organized in one file with a table of contents or bookmarks. If material pertains to a specific submission form number, that should be indicated. Requested information should be provided in the submission form and required attachments. There is no guarantee that supplemental materials will be reviewed.

Available at measure-specific web page URL identified in S.1 **Attachment:**

Contact Information

Co.1 **Measure Steward (Intellectual Property Owner):** The Joint Commission
Co.2 **Point of Contact:** Ann, Watt, awatt@jointcommission.org, 630-792-5944-
Co.3 **Measure Developer if different from Measure Steward:** The Joint Commission

Co.4 **Point of Contact:** Ann, Watt, awatt@jointcommission.org, 630-792-5944-

Additional Information

Ad.1 **Workgroup/Expert Panel involved in measure development**

Provide a list of sponsoring organizations and workgroup/panel members' names and organizations. Describe the members' role in measure development.

Michael Ross, MD, MPH (Chair)
Harbor-UCLA Medical Center
Torrance, CA

Wanda Barfield, MD, MPH
Centers for Disease Control and Prevention
Atlanta, GA

Kenneth E. Brown, MD, MBA, FACOG, FACHE
Woman´s Hospital
Lafayette, LA

Martin McCaffrey, MD
UNC North Carolina Children's Hospital
Chapel Hill, NC

Cathy Collins-Fulea, MSN, CNM
Henry Ford Hospital
Detroit, MI

Janet H. Muri, MBA
National Perinatal Information Center/
Quality Analytic Services
Providence, RI

Kathleen Simpson, PhD, RNC, FAAN
St. John's Mercy Medical Center
St. Louis, MO

Michael Socol, MD
Northwestern University Medical School
Chicago, IL

Rebecca Zimmermann, MPP
America's Health Insurance Plans
Washington, DC

The technical advisory panel (TAP) members determined priority areas that could be evaluated to improve care related to perinatal care during the development timeframe. After implementation, minor revisions, acknowledged by TAP representatives, were made to improve clarity. Hospital feedback will be reviewed during the reliability testing phase of the project to assist the TAP in making the final measure recommendations.

Measure Developer/Steward Updates and Ongoing Maintenance

Ad.2 **Year the measure was first released:** 2010
Ad.3 **Month and Year of most recent revision:** 10, 2015
Ad.4 **What is your frequency for review/update of this measure?** Biannual
Ad.5 **When is the next scheduled review/update for this measure?** 02, 2016

associated documentation based on the published manual production timelines. Ad.7 **Disclaimers:**
Ad.8 **Additional Information/Comments:**

MEASURE WORKSHEET

This document summarizes the evaluation of the measure as it progresses through NQF's Consensus Development Process (CDP). The information submitted by measure developers/stewards is included after the Brief Measure Information, Preliminary Analysis, and Pre-meeting Public and Member Comments sections.

To navigate the links in the worksheet: Ctrl + click link to go to the link; ALT + LEFT ARROW to return

Brief Measure Information

NQF #: 0475
Measure Title: Hepatitis B Vaccine Coverage Among All Live Newborn Infants Prior to Hospital or Birthing Facility Discharge
Measure Steward: Centers for Disease Control and Prevention
Brief Description of Measure: Percent of live newborn infants that receive Hepatitis B vaccination before discharge (or within 1 month of life, if the infant had an extended hospital stay) at each single hospital/birthing facility during given time period (one year).
Developer Rationale: Prevention of chronic Hepatitis B cases and its morbidity and mortality will be strengthened through the safety net provided by administration of the birth dose of the Hepatitis B vaccine to all infants before hospital or birthing facility discharge. The measure highlights the critical importance of the birth dose of Hepatitis B vaccine as a safety net for all infants. It provides an incentive to hospitals/birthing facilities to establish policies and address barriers to ensure a Hepatitis B birth dose for all infants born to consenting parents.

Numerator Statement: The number of live newborn infants administered Hepatitis B vaccine prior to discharge (or within 1 month of life, if the infant had an extended hospital stay) from the hospital/birthing facility ("birth dose" of Hepatitis B vaccine).
Denominator Statement: The number of live newborn infants born at the hospital/birthing facility during the reporting window (one calendar year).
Denominator Exclusions: a. Determine number of live newborn infants born at the hospital/birthing facility whose parent/guardian refused Hepatitis B birth dose and exclude from the denominator. ICD-10 code for this information will include the following (link: http://www.icd10data.com/ICD10CM/Codes/Z00-Z99/Z20-Z28/Z28-/#Z28):
i. Z28.82 Immunization not carried out because of caregiver refusal

Measure Type: Process
Data Source: Electronic Clinical Data, Electronic Clinical Data : Electronic Health Record, Electronic Clinical Data : Pharmacy, Electronic Clinical Data : Registry, Other, Paper Medical Records
Level of Analysis: Facility

IF Endorsement Maintenance – Original Endorsement Date: Oct 24, 2008 **Most Recent Endorsement Date:** Mar 30, 2012

Maintenance of Endorsement -- Preliminary Analysis

To maintain NQF endorsement endorsed measures are evaluated periodically to ensure that the measures still meets the NQF endorsement criteria ("maintenance"). The emphasis for maintaining endorsement is focused on how effective the measure is for promoting improvements in quality. Endorsed measures should have some experience from the field to inform the evaluation. The emphasis for maintaining endorsement is noted for each criterion.

Criteria 1: Importance to Measure and Report

1a. Evidence
Maintenance measures – less emphasis on evidence unless there is new information or change in evidence since the prior evaluation.

1a. Evidence. The evidence requirements for a *process or intermediate outcome* measure is that it is based on a systematic review (SR) and grading of the body of empirical evidence where the specific focus of the evidence matches what is being measured.

The developer provides the following evidence for this measure:

- **Systematic Review of the evidence specific to this measure?** ☒ **Yes** ☐ **No**
- **Quality, Quantity and Consistency of evidence provided?** ☒ **Yes** ☐ **No**
- **Evidence graded?** ☒ **Yes** ☐ **No**

Summary of prior review in 2012

The evidence supporting this measure are 1) guidelines from the (ACIP) Part 1: Immunization of Infants, Children, and Adolescents Report (2005) and three systematic reviews. The systematic reviews agreed that the hepatitis B vaccine, given with 24 hours of birth prevents infect, the frequency of and death as a result hepatocellular carcinoma. The guidelines also explain that if taken within 7 days, the hepatitis B vaccine prevents infection. The ACPI guideline was not graded but the three systematic reviews were graded at moderate or low. Moderate=Moderately confident in the estimate of effect on health outcome. The true effect is likely to be close to the estimate of the effect (level 3) Low=Confidence in the estimate of the effect on the health outcome is limited (level 2)

Changes to evidence from last review

☐ **The developer attests that there have been no changes in the evidence since the measure was last evaluated.**

☒ **The developer provided updated evidence for this measure:**

Updates: The developer added four systematic reviews. These sources agree and demonstrate that hepatitis B vaccine administered shortly effectively prevents perinatal hepatitis B transmission.

Exception to evidence: NA

Guidance from the Evidence Algorithm

Process measure (Box 1) → Systematic review (Box 3) → QQC present (Box 4) →recent SRs agree with benefit of neonatal hepatitis B vaccination → moderate or high

Questions for the Committee:

- *Although new studies have been provided, the underlying evidence presented appears to be the directionally the same/stronger since the last NQF endorsement review. Does the Committee agree and so there is no need for repeat discussion and vote on Evidence?*

Preliminary rating for evidence: ☒ **Pass** ☐ **No Pass**

1b. Gap in Care/Opportunity for Improvement and 1b. disparities
Maintenance measures – increased emphasis on gap and variation

1b. Performance Gap. The performance gap requirements include demonstrating quality problems and opportunity for improvement.

The developer provides several sources of data on performance:

- The 2014 National Immunization Survey reported Hepatitis B birth dose rates by state among infants born from January 2011-May 2013: mean coverage of 72.4%±1.5, which was down from 74.2%±1.4 in 2013. Among U.S. Department of Health and Human Services states and local areas, the minimum coverage was 48.4%±7.5 and the maximum coverage was 88.4%±4.6.
- 40 delivery facilities in New York City (118,995 total births during 2013): 67.4% of infants received the birth dose

of Hepatitis B vaccine. This percentage is unchanged from 2012. Birth dose coverage ranged from a low of 5.4% at one facility to a high of 102.7% (note: 102.7% is the reported figure) at another facility. [c]

- 14 hospitals in New York City and Michigan (36,046 live births during 2013): 77.3% of infants received the birth dose of Hepatitis B vaccine; this percentage represents an increase from 72.6% of infants during 2012.
- 119 hospitals in Texas during 2009-2010 demonstrated the 0-3 day of life Hepatitis B vaccination rates by birthing facility: mean 90.4%; median 95.5%; maximum 100%; minimum 21.2%.

Disparities

The developer provides literature references rather than data from use of this measure:

- A 2015 review of Perinatal exposure to hepatitis B found that between 2007 to 2013 Perinatal transmission occurred in 1.4% of infants born to mothers whose race was Asian/Pacific Islander, compared to 0.5%, 0.1%, and 0.6% of mothers whose race/ethnicity was black, white, or Hispanic, respectively ($p<0.01$). Perinatal transmission occurred in 1.3% of infants whose mothers were foreign-born, compared to 0.6% of infants whose mothers were U.S.-born ($p=0.02$).
- A 2006 study 'Estimating the Number of Births to Hepatitis B Virus-infected Women in 22 States' found that that foreign-born women from countries highly endemic for Hepatitis B infection (despite being a minority of all women giving birth), U.S.- and Canadian-born non-Hispanic blacks, and Asian/Pacific Islanders represented the majority of all births to HBsAg-positive women. Of 2,359,912 births in the 22 states evaluated, approximately 16,500 births were estimated to be from HBV-infected women; 80.6% of these were foreign-born women.

Questions for the Committee:

- *Is there a gap in care that warrants a national performance measure?*
- *Does this measure provide information to understand disparities in this area of healthcare?*

Preliminary rating for opportunity for improvement: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 1: Importance to Measure and Report (including 1a, 1b, 1c)

1a. Evidence to Support Measure Focus

Comments:

1b. Performance Gap

Comments:

Significant performance gap noted.

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability

2a1. Reliability Specifications

Maintenance measures – no change in emphasis – specifications should be evaluated the same as with new measures

2a1. Specifications requires the measure, as specified, to produce consistent (reliable) and credible (valid) results about the quality of care when implemented.

Data source(s):

- Electronic Clinical Data; Electronic Health Records, Pharmacy, Registry; Paper Medical Records

Specifications:

- The level of analysis is at the facility and the care setting is hospital/acute care facility.
- The numerator is the number of live newborn infants administered Hepatitis B vaccine prior to discharge (or within 1 month of life, if the infant had an extended hospital stay) from the hospital/birthing facility ("birth dose" of Hepatitis B vaccine).
- The denominator includes the number of live newborn infants born at the hospital/birthing facility during the

reporting window which is one calendar year.

- The denominator exclusions include the number of infants born at the facility during one calendar year whose parent/guardian refused administration of a birth dose of Hepatitis B vaccine before discharge (or by 1 month of age if during a prolonged stay) from the hospital/birthing facility.
- An attached spreadsheet contains numerous ICD-9 and ICD-10 codes for the outcome of the delivery, live birth infant type of birth, and reason for why the vaccination was not carried out.
- A calculation algorithm describes the process of calculating the measure.

Questions for the Committee:

- *Are all the data elements clearly defined? Are all appropriate codes included?*
- *Is the logic or calculation algorithm clear?*
- *Is it likely this measure can be consistently implemented?*
- *Is refusal of treatment an appropriate reason for exclusion from the denominator?*

2a2. Reliability Testing Testing attachment
Maintenance measures – less emphasis if no new testing data provided

2a2. Reliability testing demonstrates if the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise enough to distinguish differences in performance across providers.

For maintenance measures, summarize the reliability testing from the prior review:

For the prior review the developers presented results of feasibility testing. Reliability was evaluated by the level of concordance between the rates calculated by the facilities in the Feasibility Study when compared with the rates determined by medical chart review in the PHEP. Results showed a variance of ± 10% for most hospitals that provided a result (30 of 36 hospitals).

Describe any updates to testing – The developers have presented new testing of the measure score using signal-to-noise method -described in RED.

SUMMARY OF TESTING

Reliability testing level ☒ **Measure score** ☐ **Data element** ☐ **Both**

Reliability testing performed with the data source and level of analysis indicated for this measure ☐ **Yes** ☒ **No**

Method of reliability testing

- Signal to noise analysis was performed over 4 years 2013 as described below. Signal to noise is an appropriate method of assessing differences in measure performance. Generally, a reliability of >0.70 is considered acceptable.

Results of reliability testing

Distribution of reliability statistics from signal-to-noise analysis:

Year	Mean	Median	Min	Max	Hospitals	Live Births
2013	0.99844	0.99916	0.99251	0.99999	14	32,471
2012	0.99900	0.99944	0.99631	1.00000	13	33,078
2011	0.99834	0.99900	0.99660	0.99999	11	24,914
2010	0.99368	0.99388	0.98132	1.00000	8	16,468

- The developer states that reliability between 0.98132 and 1.00000 indicate very high reliability, indicating that variability between hospitals regarding the Hepatitis B vaccine birth dose is due to actual performance differences rather than measurement error.

Questions for the Committee:

- *Is the test sample adequate to generalize for widespread implementation?*
- *Do the results demonstrate sufficient reliability so that differences in performance can be identified?*

Guidance from algorithm #2 Evaluating Reliability: Precise specifications (Box 1) → Empirical reliability testing (Box 2) → Computed performance measure scores (Box 4) → Signal-to-Noise Analysis (Box 5) → Statistic and Scope of testing (Box 6a) → Rate as high.

Preliminary rating for reliability: ☒ **High** ☐ **Moderate** ☐ **Low** ☐ **Insufficient**

2b. Validity
Maintenance measures – less emphasis if no new testing data provided

2b1. Validity: Specifications

2b1. Validity Specifications. This section should determine if the measure specifications are consistent with the evidence.

Specifications consistent with evidence in 1a. ☒ **Yes** ☐ **Somewhat** ☐ **No**
Specification not completely consistent with evidence

Question for the Committee:

- *Are the specifications consistent with the evidence?*

2b2. Validity testing

2b2. Validity Testing should demonstrate the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality.

For maintenance measures, summarize the validity testing from the prior review: face validity only

Describe any updates to validity testing: Updated face validity testing was included.

SUMMARY OF TESTING
Validity testing level ☒ **Measure score** ☐ **Data element testing against a gold standard** ☐ **Both**

Method of validity testing of the measure score:

- ☒ **Face validity only**
- ☐ **Empirical validity testing of the measure score**

Validity testing method:

- The developer used Spertus et al "American College of Cardiology and American Heart Association Methodology for the Selection and Creation of Performance Measures for Quantifying the Quality of Cardiovascular Care" method of face validity testing. Face validity was assessed systematically, in a transparent process, by recognized experts in the field to assess the computed performance measure score as a method to distinguish between good and poor quality. A rating survey was distributed to 22 experts and 14 (response rate = 63.6) responded.

Validity testing results:

- The developer provides a results of the validity survey using a 1-5 scale with the mean responses ranging from 4.07 to 4.50, 'agree' and 'strongly agree'.

Staff note: Since only face validity was assessed rather than empirical validity testing, the highest rating possible for validity is MODERATE.

Questions for the Committee:

- *Do the results demonstrate sufficient validity so that conclusions about quality can be made?*
- *Do you agree that the score from this measure as specified is an indicator of quality?*

2b3-2b7. Threats to Validity

2b3. Exclusions:

The developer did not supply the frequency of exclusions but did supply their reliability results from testing the measure without exclusions:

Year	Mean	Median	Min	Max	# Hospitals	Live births
2013	0.99738	0.99865	0.98977	0.99999	14	36,046
2012	0.99852	0.99904,	0.99446,	1.00000	13	36,790
2011	0.99923	0.99957	0.99712	0.99999	11	32,862
2010	0.99856	0.99903	0.99525	0.99986	8	23,275

*Please see reliability results *with* exclusions above

The developer's analysis, not excluding refusals from the denominator, ranges from 0.98977 to 1.00000. This indicate very high reliability, indicating that variability between hospitals regarding the Hepatitis B vaccine birth dose is due to actual performance differences rather than measurement error. The developer states that these results support consideration of removing exclusions from the denominator.

Questions for the Committee:

- *Is refusal of treatment a valid reason for exclusion from the denominator?*
- *Are the exclusions consistent with the evidence?*
- *Are any patients or patient groups inappropriately excluded from the measure?*
- *Are the exclusions/exceptions of sufficient frequency and variation across providers to be needed (and outweigh the data collection burden)?*

2b4. Risk adjustment: **Risk-adjustment method** ☒ **None** ☐ **Statistical model** ☐ **Stratification**

The developer states that risk adjustment or stratification is not warranted for this measure, as the process is universally recommended for all infants. The validity survey indicated that half (7 of 14) respondents strongly agreed that this measure addresses an area under the practitioner's control, and 35.7% (5 of 14) agreed it addresses in area under the practitioner's control.

2b5. Meaningful difference *(can statistically significant and clinically/practically meaningful differences in performance measure scores can be identified)*:

The developer reports the following descriptive analysis of the measure results:

	2013		2012		2011		2010	
	w/ Exclusions	W/out Exclusions	w/ Exclusions	W/out Exclusions	w/ Exclusions	W/out Exclusions	w/ Exclusions	W/out Exclusions
# of facilities	14	14	13	13	11	11	8	8
Total # of live births, excluding	32,471	36,046	33,078	36,790	24,914	32,862	16,468	23,275

refusals								
Min birth dose coverage	29.95%	29.95%	30.89%	30.89%	32.37%	25.61%	81.55%	33.88%
Max birth dose coverage	99.96%	99.96%	100.00%	100.00%	99.91%	99.91%	100.00%	98.67%
Mean birth dose coverage	91.09%	84.15%	87.38%	80.62%	87.41%	72.32%	91.73%	73.89%
Median birth dose coverage	96.08%	90.78%	95.47%	90.23%	96.25%	93.95%	94.19%	90.51%

Descriptive statistics were used to compare coverage between facilities. In 2013 there is variability between the 14 hospitals assessed for Hepatitis B vaccine birth dose coverage, ranging from a low of 29.95% to a high of 99.96. It is also noted that the mean birth dose coverage differs from the median coverage, indicating birth dose coverage is skewed and not normally distributed. The national Hepatitis B vaccine birth dose coverage overall was 72.4% in 2014 (National Immunization Survey). The developer states that these data demonstrate clinically meaningful differences in performance affecting a large patient population (all newborns) for which continued improvement is warranted, and could help to justify not excluding parent refusals from the denominator, thereby making the measure more robust.

Question for the Committee:

- *Does this measure identify meaningful differences about quality?*

2b6. Comparability of data sources/methods: N/A

2b7. Missing Data

- The developer states "By definition, data missing regarding administration of the Hepatitis B vaccine birth dose correspond to no birth dose administration, as 'negatives' (i.e., lack of vaccination) are not or may not be billed for, coded, or documented. "

Guidance from algorithm #3 Evaluating Validity: Specifications consistent with evidence (Box 1) → Assessed all potential threats to validity (Box 2) → testing of face validity (Box 4) → moderate (highest rating possible in the absence of empirical validity testing)

Preliminary rating for validity: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 2: Scientific Acceptability of Measure Properties (including all 2a, 2b, and 2d)

2a1. & 2b1. Specifications
Comments:

2a2. Reliability Testing
Comments:

2b2. Validity Testing
Comments:

2b3. Exclusions Analysis
2b4. Risk Adjustment/Stratification for Outcome or Resource Use Measures
2b5. Identification of Statistically Significant & Meaningful Differences In Performance
2b6. Comparability of Performance Scores When More Than One Set of Specifications

2b7. Missing Data Analysis and Minimizing Bias
Comments:
No concerns.

Criterion 3. Feasibility
Maintenance measures – no change in emphasis – implementation issues may be more prominent

3. Feasibility is the extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

- This measure is based on electronic health records, electronic clinical pharmacy data, electronic registry data, and paper medical records.
- The developer notes that, "not all hospitals/birthing facilities currently have the infrastructure for electronic sources, therefore paper sources of medical records, pharmacy, etc. must be used until such time as all hospital/birthing facilities have electronic source capabilities. However, adoption of electronic health records has increased; in 2014, 75.5% of acute care hospitals had adopted at least a basic electronic health record, representing an eight-fold increase from 2008." It is also noted that the data for the measure is coded and abstracted by someone other than the person who originally obtained the information.

Questions for the Committee:
- *Are the required data elements routinely generated and used during care delivery?*
- *Are the required data elements available in electronic form, e.g., EHR or other electronic sources?*
- *Is the data collection strategy ready to be put into operational use?*

Preliminary rating for feasibility: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments
Criteria 3: Feasibility

3a. Byproduct of Care Processes
3b. Electronic Sources
3c. Data Collection Strategy
Comments:
No concerns.

Criterion 4: Usability and Use
Maintenance measures – increased emphasis – much greater focus on measure use and usefulness, including both impact /improvement and unintended consequences

4. Usability and Use evaluate the extent to which audiences (e.g., consumers, purchasers, providers, policymakers) use or could use performance results for both accountability and performance improvement activities.

Current uses of the measure
Publicly reported? ☒ **Yes** ☐ **No**

Current use in an accountability program? ☒ **Yes** ☐ **No**
OR
Planned use in an accountability program? ☐ **Yes** ☐ **No**

Accountability program details
- New York City Department of Health and Mental Hygiene http://www.nyc.gov/html/doh/html/hcp/cd-hepatitisb-pregnancy.shtml

Improvement results

- Data is provided from the 'National Immunization Survey' which includes data from 14,893 children, which demonstrates that Hepatitis B birth dose coverage has improved from 64.1% (+/-1.3) in 2010 to 72.4% (+/-1.5) in 2014.

Unexpected findings (positive or negative) during implementation

Potential harms

- The developer notes that an unintended consequence of reporting the measure after removing parent/guardian refusals will be the loss of information about hospitals with an unusually high refusal rates.
- It is noted that hospitals that do not have the means to capture electronic data face more measurement burden as it takes more time and effort to collect the required information than hospitals that are able to capture electronic data.

Feedback:

***Questions for the Committee**:*

- *How can the performance results be used to further the goal of high-quality, efficient healthcare?*
- *Do the benefits of the measure outweigh any potential unintended consequences?*

Preliminary rating for usability and use: ☒ **High** ☐ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 4: Usability and Use

4a. Accountability and Transparency
4b. Improvement
4c. Unintended Consequences
Comments:

Criterion 5: Related and Competing Measures

Related or competing measures

- none

Harmonization

- N/A

Pre-meeting public and member comments

-

NATIONAL QUALITY FORUM—Evidence (subcriterion 1a)

Measure Number (*if previously endorsed*)**:** 0475
Measure Title: Hepatitis B Vaccine Coverage Among All Live Newborn Infants Prior to Hospital or Birthing Facility Discharge
IF the measure is a component in a composite performance measure, provide the title of the Composite Measure here: N/A

Date of Submission: 2/16/2016

Instructions

- *For composite performance measures:*
  - *A separate evidence form is required for each component measure unless several components were studied together.*
  - *If a component measure is submitted as an individual performance measure, attach the evidence form to the individual measure submission.*
- Respond to all questions as instructed with answers immediately following the question. All information needed to demonstrate meeting the evidence subcriterion (1a) must be in this form. An appendix of *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 10 pages (*incudes questions/instructions*; minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.

Note: The information provided in this form is intended to aid the Steering Committee and other stakeholders in understanding to what degree the evidence for this measure meets NQF's evaluation criteria.

1a. Evidence to Support the Measure Focus
The measure focus is evidence-based, demonstrated as follows:

- Health outcome: [3] a rationale supports the relationship of the health outcome to processes or structures of care. Applies to patient-reported outcomes (PRO), including health-related quality of life/functional status, symptom/symptom burden, experience with care, health-related behavior.
- Intermediate clinical outcome: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured intermediate clinical outcome leads to a desired health outcome.
- Process: [5] a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured process leads to a desired health outcome.
- Structure: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured structure leads to a desired health outcome.
- Efficiency: [6] evidence not required for the resource use component.

Notes

3. Generally, rare event outcomes do not provide adequate information for improvement or discrimination; however, serious reportable events that are compared to zero are appropriate outcomes for public reporting and quality improvement.
4. The preferred systems for grading the evidence are the U.S. Preventive Services Task Force (USPSTF) grading definitions and methods, or Grading of Recommendations, Assessment, Development and Evaluation (GRADE) guidelines.
5. Clinical care processes typically include multiple steps: assess → identify problem/potential problem → choose/plan intervention (with patient input) → provide intervention → evaluate impact on health status. If the measure focus is one step in such a multistep process, the step with the strongest evidence for the link to the desired outcome should be selected as the focus of measurement. Note: A measure focused only on collecting PROM data is not a PRO-PM.
6. Measures of efficiency combine the concepts of resource use and quality (see NQF's Measurement Framework: Evaluating Efficiency Across Episodes of Care; AQA Principles of Efficiency Measures).

1a.1.This is a measure of: (*should be consistent with type of measure entered in De.1*)
Outcome

☐ Health outcome: Click here to name the health outcome

☐Patient-reported outcome (PRO): Click here to name the PRO
PROs include HRQoL/functional status, symptom/symptom burden, experience with care, health-related behaviors
☐ Intermediate clinical outcome (*e.g., lab value*): Click here to name the intermediate outcome
☒ Process: Newborn Hepatitis B vaccination
☐ Structure: Click here to name the structure
☐ Other: Click here to name what is being measured

HEALTH OUTCOME/PRO PERFORMANCE MEASURE *If not a health outcome or PRO, skip to 1a.3*

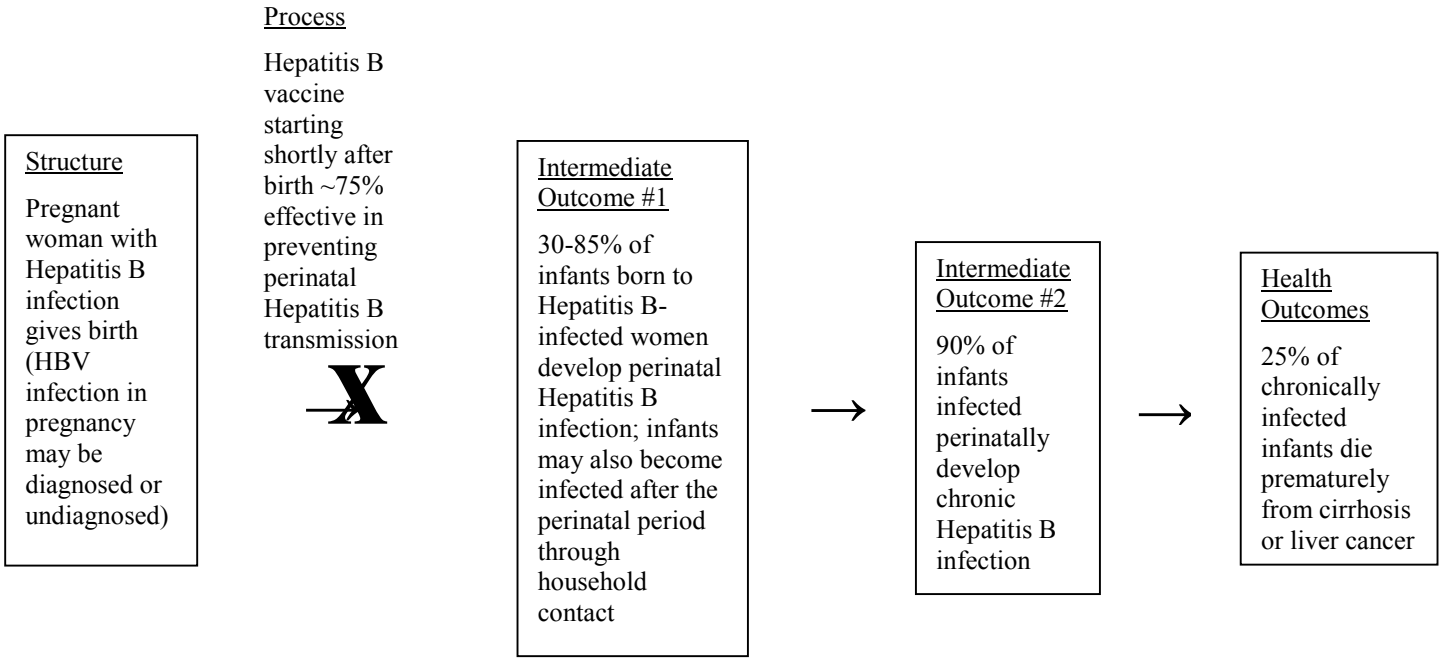
1a.2. Briefly state or diagram the path between the health outcome (or PRO) and the healthcare structures, processes, interventions, or services that influence it.

1a.2.1. State the rationale supporting the relationship between the health outcome (or PRO) to at least one healthcare structure, process, intervention, or service (*i.e., influence on outcome/PRO*).

Note: For health outcome/PRO performance measures, no further information is required; however, you may provide evidence for any of the structures, processes, interventions, or service identified above.

INTERMEDIATE OUTCOME, PROCESS, OR STRUCTURE PERFORMANCE MEASURE

1a.3. Briefly state or diagram the path between structure, process, intermediate outcome, and health outcomes. Include all the steps between the measure focus and the health outcome.



1a.3.1. What is the source of the systematic review of the body of evidence that supports the performance measure?

☒ Clinical Practice Guideline recommendation – ***complete sections 1a.4, and 1a.7***
☐ US Preventive Services Task Force Recommendation – ***complete sections 1a.5 and 1a.7***
☒ Other systematic review and grading of the body of evidence (*e.g., Cochrane Collaboration, AHRQ Evidence Practice Center*) – ***complete sections 1a.6 and 1a.7***
☐ Other – ***complete section 1a.8***

Please complete the sections indicated above for the source of evidence. You may skip the sections that do not apply.

1a.4. CLINICAL PRACTICE GUIDELINE RECOMMENDATION

1a.4.1. Guideline citation (*including date*) and **URL for guideline** (*if available online*):

A Comprehensive Immunization Strategy to Eliminate Transmission of Hepatitis B Virus Infection in the United States. Recommendations of the Advisory Committee on Immunization Practices (ACIP) Part 1: Immunization of Infants, Children, and Adolescents. Prepared by Eric E. Mast, Harold S. Margolis, Anthony E. Fiore, Edward W. Brink, Susan T. Goldstein, Susan A. Wang, Linda A. Moyer, Beth P. Bell, Miriam J. Alter. MMWR Recomm Rep. 2005 Dec 23;54(RR16):1-31
http://www.cdc.gov/mmwr/preview/mmwrhtml/rr5416a1.htm, accessed 12/17/2015

1a.4.2. Identify guideline recommendation number and/or page number and **quote verbatim, the specific guideline recommendation**.

Pages 15 and 17 (Page 16 is a Box figure)

"•All infants should receive the hepatitis B vaccine series as part of the recommended childhood immunization schedule (Table 5 and Appendix B). (For recommendations on management of infants born to HBsAg-positive mothers and infants born to mothers with unknown HBsAg status, see Prevention of Perinatal HBV Infection and Management of Pregnant Women.)
•For all medically stable infants weighing >2,000 g at birth and born to HBsAg-negative mothers, the first dose of vaccine should be administered before hospital discharge. Only single-antigen hepatitis B vaccine should be used for the birth dose.
•On a case-by-case basis and only in rare circumstances, the first dose may be delayed until after hospital discharge for an infant who weighs >2,000 g and whose mother is HBsAg negative.
--- When such a decision is made, a physician's order to withhold the birth dose and a copy of the original laboratory report indicating that the mother was HBsAg negative during this pregnancy should be placed in the infant's medical record.
--- For infants who do not receive a first dose before hospital discharge, the first dose should be administered no later than age 2 months.
--- Situations in which the birth dose should not be delayed include any high-risk sexual or drug-using practices of the infant's mother during pregnancy (e.g., having had more than one sex partner during the previous 6 months or an HBsAg-positive sex partner, evaluation or treatment for an STD, or recent or current injection-drug use) and expected poor compliance with follow-up to initiate the vaccine series.
•Preterm infants weighing <2,000 g and born to HBsAg-negative mothers should have their first vaccine dose delayed until 1 month after birth or hospital discharge (Table 4). For these infants, a copy of the original laboratory report indicating that the mother was HBsAg negative during this pregnancy should be placed in the infant's medical record."

1a.4.3. Grade assigned to the quoted recommendation with definition of the grade:

N/A (At the time the ACIP Guideline was developed, ACIP did not use a grading of evidence approach. However, ACIP recommends the Hepatitis B vaccine birth dose universally and unequivocally for all infants, prior to hospital discharge).

1a.4.4. Provide all other grades and associated definitions for recommendations in the grading system. (*Note: If separate grades for the strength of the evidence, report them in section 1a.7.*)

N/A

1a.4.5. Citation and URL for methodology for grading recommendations (*if different from 1a.4.1*)**:**

N/A

1a.4.6. If guideline is evidence-based (rather than expert opinion), are the details of the quantity, quality, and consistency of the body of evidence available (e.g., evidence tables)?

☐ Yes → ***complete section 1a.7***

☐ No → ***report on another systematic review of the evidence in sections 1a.6 and 1a.7; if another review does not exist, provide what is known from the guideline review of evidence in 1a.7***

N/A

1a.5. UNITED STATES PREVENTIVE SERVICES TASK FORCE RECOMMENDATION

1a.5.1. Recommendation citation (*including date*) and **URL for recommendation** (*if available online*):

N/A

1a.5.2. Identify recommendation number and/or page number and **quote verbatim, the specific recommendation**.

N/A

1a.5.3. Grade assigned to the quoted recommendation with definition of the grade:

N/A

1a.5.4. Provide all other grades and associated definitions for recommendations in the grading system. (*Note: the grading system for the evidence should be reported in section 1a.7.*)

N/A

1a.5.5. Citation and URL for methodology for grading recommendations (*if different from 1a.5.1*)**:**
N/A

Complete section 1a.7

1a.6. OTHER SYSTEMATIC REVIEW OF THE BODY OF EVIDENCE

1a.6.1. Citation (*including date*) and **URL** (*if available online*):

Weekly Epidemiological Record. Hepatitis B Vaccine: WHO Position Paper. World Health Organization. 2009;84,405-420
http://www.who.int/wer/2009/wer8440.pdf?ua=1, accessed 12/17/2015

Strategic Advisory Group of Experts (SAGE) Hepatitis Working Group, Hepatitis B Grading of Scientific Evidence (Hepatitis B vaccine within 24 hours of birth)
http://www.who.int/immunization/hepb_grad_24hours.pdf?ua=1, accessed 12/16/2015

Strategic Advisory Group of Experts (SAGE) Hepatitis Working Group, Hepatitis B Grading of Scientific Evidence (Hepatitis B vaccine within 7 days of birth)
http://www.who.int/immunization/hepb_grad_7days.pdf?ua=1, accessed 12/16/2015

1a.6.2. Citation and URL for methodology for evidence review and grading (*if different from 1a.6.1*)**:**
Please refer to 1a.6.1

Guidance for the Development of Evidence-Based Vaccine-Related Recommendations, Version 5. World Health Organization. 2015 Oct 8
http://www.who.int/immunization/sage/Guidelines_development_recommendations.pdf?ua=1, accessed 12/17/2015

Complete section 1a.7

_____

1a.7. FINDINGS FROM SYSTEMATIC REVIEW OF BODY OF THE EVIDENCE SUPPORTING THE MEASURE

If more than one systematic review of the evidence is identified above, you may choose to summarize the one (or more) for which the best information is available to provide a summary of the quantity, quality, and consistency of the body of evidence. Be sure to identify which review is the basis of the responses in this section and if more than one, provide a separate response for each review.

1a.7.1. What was the specific structure, treatment, intervention, service, or intermediate outcome addressed in the evidence review?

Hepatitis B vaccine with 24 hours of birth to prevent Hepatitis B infection, incidence of hepatocellular carcinoma, and mortality from hepatocellular carcinoma

Hepatitis B vaccine with 7 days of birth to prevent Hepatitis B infection and chronic Hepatitis B infection

1a.7.2. Grade assigned for the quality of the quoted evidence with definition of the grade:

-Moderate quality evidence to support effectiveness of Hepatitis B vaccine given within 24 hours of birth to prevent Hepatitis B infection
-Low quality evidence to support effectiveness of Hepatitis B vaccine given within 24 hours of birth to prevent incidence of hepatocellular carcinoma
-Low quality evidence to support effectiveness of Hepatitis B vaccine given within 24 hours of birth to prevent mortality from hepatocellular carcinoma.

-Moderate quality evidence to support effectiveness of Hepatitis B vaccine given within 7 days of birth to prevent Hepatitis B infection
-Moderate quality evidence to support effectiveness of Hepatitis B given within 7 days of birth to prevent chronic Hepatitis B infection

Moderate=Moderately confident in the estimate of effect on health outcome. The true effect is likely to be close to the estimate of the effect (level 3)
Low=Confidence in the estimate of the effect on the health outcome is limited (level 2)

1a.7.3. Provide all other grades and associated definitions for strength of the evidence in the grading system.

High=Very confident that the true effect lies close to that of the estimate of effect on health outcome (level 4)
Moderate=Moderately confident in the estimate of effect on health outcome. The true effect is likely to be close to the estimate of the effect (level 3)
Low=Confidence in the estimate of the effect on the health outcome is limited (level 2)
Very low=Very little confidence in the estimate of the effect on the health outcome (level 1)

1a.7.4. What is the time period covered by the body of evidence? (*provide the date range, e.g., 1990-2010*). Date range: Click here to enter date range

1984-1997

QUANTITY AND QUALITY OF BODY OF EVIDENCE

1a.7.5. How many and what type of study designs are included in the body of evidence? (*e.g., 3 randomized controlled trials and 1 observational study*)

Hepatitis B vaccine given within 24 hours of birth
To prevent Hepatitis B infection: 5 randomized controlled trials
Incidence of hepatocellular carcinoma: 1 observational study
Mortality from hepatocellular carcinoma: 1 observational study

Hepatitis B vaccine given within 7 days of birth
To prevent Hepatitis B infection: 1 observational study
To prevent chronic Hepatitis B infection: 1 observational study

1a.7.6. What is the overall quality of evidence across studies in the body of evidence? (*discuss the certainty or confidence in the estimates of effect particularly in relation to study factors such as design flaws, imprecision due to small numbers, indirectness of studies to the measure focus or target population*)

Hepatitis B vaccine given within 24 hours of birth
To prevent Hepatitis B infection: Serious Limitations: Randomization generation unclear in all trials, allocation concealment unclear in 3 trials, losses to follow-up not described in all trials. No serious inconsistency, no serious indirectness (although 4 trials used plasma vaccine [no longer available] and 1 trial used recombinant vaccine), no serious imprecision.
Incidence of hepatocellular carcinoma: No serious limitations, no serious inconsistency, no serious indirectness, no serious imprecision.
Mortality from hepatocellular carcinoma: No serious limitations, no serious inconsistency, no serious indirectness, no serious imprecision.

Hepatitis B vaccine given within 7 days of birth
To prevent Hepatitis B infection: No serious limitations, no serious inconsistency, no serious indirectness, no serious imprecision.
To prevent chronic Hepatitis B infection: No serious limitations, no serious inconsistency, no serious indirectness, no serious imprecision.

ESTIMATES OF BENEFIT AND CONSISTENCY ACROSS STUDIES IN BODY OF EVIDENCE

1a.7.7. What are the estimates of benefit—magnitude and direction of effect on outcome(s) across studies in the body of evidence? (*e.g., ranges of percentages or odds ratios for improvement/ decline across studies, results of meta-analysis, and statistical significance*)

Hepatitis B vaccine given within 24 hours of birth
To prevent Hepatitis B infection: Number of patients vaccine versus control: 33/252 (13.1%) vs 77/151 (51%). Relative risk 0.28, 95% CI 0.2-0.4.

Hepatitis B vaccine given within 7 days of birth
To prevent Hepatitis B infection: Study showed a strong relationship between time of first dose and probability of infection (odds ratio 4.3, 95% CI 2.2-8.4) for each unit increase in age. This was statistically significant for children receiving their first dose after 7 days of age.
To prevent chronic Hepatitis B infection: Study showed a strong relationship between time of first dose and chronic infection (odds ratio 3.3, 95% CI 1.3-8.2) for each unit increase in age.

1a.7.8. What harms were studied and how do they affect the net benefit (benefits over harms)?

Of the 10 manuscripts contributing to the evidence review, 6 manuscripts did not comment regarding adverse events, 1 manuscript was written in Chinese, and 3 manuscripts documented no serious vaccine-related adverse events. Net benefits exceed net harms.

UPDATE TO THE SYSTEMATIC REVIEW(S) OF THE BODY OF EVIDENCE

1a.7.9. If new studies have been conducted since the systematic review of the body of evidence, provide for <u>each</u> new study: 1) citation, 2) description, 3) results, 4) impact on conclusions of systematic review.

Hepatitis B vaccine efficacy in infants is well-established, and, as such, additional randomized controlled trials to assess efficacy would be considered unethical. However, other studies which support the Hepatitis B vaccine birth dose are summarized here:

Outcomes of Infants Born to Women Infected with Hepatitis B. S. Schillie, T. Walker, S. Veselsky, S. Crowley, C. Dusek, J. Lazaroff, S. Morris, K. Onye, S. Ko, N. Fenlon, N. Nelson, T. Murphy. Pediatrics. 2015 May;135:e1141-7.

Data from 5 U.S. Perinatal Hepatitis B Prevention Programs during 2007-2013 were analyzed to assess outcomes of infants born to Hepatitis B-infected mothers. Hepatitis B vaccine was administered within 12 hours of birth to 96.4% of infants. Among 9,252 infants with information, 1.1% developed infection. The authors conclude that timely infant immunoprophylaxis and completion of the Hepatitis B vaccine series prevents most perinatal Hepatitis B transmission. The findings from the study support the evidence that Hepatitis B vaccine shortly after birth prevents perinatal Hepatitis B infection.

Effects of Hepatitis B Immunization on Prevention of Mother-to-Infant Transmission of Hepatitis B Virus and on the Immune Response of Infants towards Hepatitis B vaccine. L. Zhang, X. Gui, C. Teter, H. Zhong, Z. Pang, L. Ding, F. Li, Y. Zhou, L. Zhang. Vaccine. 2014 Oct 21;32:6091-7.

Infants of Hepatitis B-infected (and uninfected) mothers were studied from 15 centers in China during 2008-2013 to evaluate the effect of Hepatitis B vaccine on perinatal Hepatitis B transmission. Among infants of 1,202 mothers, 40 were infected, for an immunoprophylaxis failure rate of 3.3%. No perinatal transmission occurred among infants born to mothers who were Hepatitis B e antigen negative who received Hepatitis B vaccine at birth, regardless of Hepatitis B Immune Globulin administration. The authors conclude that Hepatitis B vaccine and Hepatitis B Immune Globulin effectively prevent perinatal Hepatitis B transmission. The findings from the study support the evidence that Hepatitis B vaccine shortly after birth prevents perinatal Hepatitis B infection.

Hepatitis B Vaccine Response among Infants Born to Hepatitis B Surface Antigen-Positive Women. S. Ko, S. Schillie, T. Walker, S. Veselsky, N. Nelson, J. Lazaroff, S. Crowley, C. Dusek, K. Loggins, K. Onye, N. Fenlon, T. Murphy. Vaccine. 2014 Apr 11;32:2127-33.

Data from uninfected infants born to infected mothers from 5 U.S. Perinatal Hepatitis B Prevention Programs during 2008-2013 were analyzed to assess infant vaccine response. The majority (94.7%) of 8,654 infants developed antibody to Hepatitis B surface antigen greater than or equal to 10 mIU/mL. The authors conclude that overall response to Hepatitis B vaccine is high among infants born to infected mothers. The findings from the study support the evidence that Hepatitis B vaccine shortly after birth prevents perinatal Hepatitis B infection.

Seroprotection after Recombinant Hepatitis B Vaccination among Newborn Infants: A Review. S. Schillie, T. Murphy. Vaccine. 2013 May 17;21:2506-16.

Studies published between 1987-2011 were reviewed to assess anti-HBs levels (a serologic correlate of protection) following infant vaccination (with recombinant vaccine) starting at birth, by maternal Hepatitis B infection status, Hepatitis B Immune Globulin administration, infant birth weight, infant gestational age, and vaccine dose and schedule. Overall, 98% of infants achieved seroprotective anti-HBs levels following a complete series. The authors conclude that high levels of protection from Hepatitis B vaccine starting at birth are achieved among term infants. The findings from the study support the evidence that Hepatitis B vaccine shortly after birth prevents perinatal Hepatitis B infection.

Hepatitis B and the Need for a Booster Dose. E. Leuridan, P. Van Damme. Clin Infect Dis. 2011 Jul 1;53:68-75.

The literature regarding Hepatitis B vaccine long-term protection (until 2010) was examined in the context of four mechanisms for assessing long-term protection: anamnestic response to a booster dose, number of infections, T- and B-cell activity, and seroepidemiology. The authors conclude that Hepatitis B vaccine protection lasts at least 20 years. The findings from this study supplement the conclusions from the evidence review.

_____

1a.8 OTHER SOURCE OF EVIDENCE

If source of evidence is NOT from a clinical practice guideline, USPSTF, or systematic review, please describe the evidence on which you are basing the performance measure.

1a.8.1 What process was used to identify the evidence?

N/A

1a.8.2. Provide the citation and summary for each piece of evidence.

N/A

1. Evidence, Performance Gap, Priority – Importance to Measure and Report

Extent to which the specific measure focus is evidence-based, important to making significant gains in healthcare quality, and improving health outcomes for a specific high-priority (high-impact) aspect of healthcare where there is variation in or overall less-than-optimal performance. ***Measures must be judged to meet all subcriteria to pass this criterion and be evaluated against the remaining criteria.***

1a. Evidence to Support the Measure Focus – See attached Evidence Submission Form
Template_MeasSubm_Evidence.docx

1b. Performance Gap

Demonstration of quality problems and opportunity for improvement, i.e., data demonstrating:

- considerable variation, or overall less-than-optimal performance, in the quality of care across providers; and/or
- disparities in care across population groups.

1b.1. Briefly explain the rationale for this measure *(e.g., the benefits or improvements in quality envisioned by use of this measure)*
Prevention of chronic Hepatitis B cases and its morbidity and mortality will be strengthened through the safety net provided by administration of the birth dose of the Hepatitis B vaccine to all infants before hospital or birthing facility discharge. The measure highlights the critical importance of the birth dose of Hepatitis B vaccine as a safety net for all infants. It provides an incentive to hospitals/birthing facilities to establish policies and address barriers to ensure a Hepatitis B birth dose for all infants born to consenting parents.

1b.2. Provide performance scores on the measure as specified (current and over time) at the specified level of analysis. *(This is required for endorsement maintenance. Include mean, std dev, min, max, interquartile range, scores by decile. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included). This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*
a. NQF MEASURE Feasibility Study: The Feasibility Study was conducted in 50 hospitals using 2008 data that were representative of labor and delivery hospitals in Texas. The annual birth cohort size at study hospitals ranged from 219 to 6,530 births and together represented more than 100,000 births. Hospitals were divided into urban and rural locations and included for-profit, not-for-profit, and public facilities. Thirty-six of 50 hospitals (72%), representing more than 62,000 births, calculated the measure. The Hepatitis B birth dose vaccination rates not excluding parent/guardian refusals were: median 94.5%; minimum 8%; maximum 100%.[a]
b. Estimates from a variety of other sources demonstrate wide variation in performance across providers, and substantial room for improvement.
i. The 2014 National Immunization Survey reported Hepatitis B birth dose rates by state and jurisdiction using verified vaccination records for 0-3 days of life among infants born from January 2011-May 2013 (the latest available). The survey reported a mean coverage of 72.4%±1.5, which was down from 74.2%±1.4 in 2013. Among U.S. Department of Health and Human Services states and local areas, the minimum coverage was 48.4%±7.5 and the maximum coverage was 88.4%±4.6.[b]
ii. Among 40 delivery facilities in New York City, representing 118,995 total births during 2013, 67.4% of infants received the birth dose of Hepatitis B vaccine, either given at day 0 (birth), day 1, day 2, and day 3. (Note that infants with birth weights less than 2000 grams were included in this calculation). This percentage is unchanged from 2012. Birth dose coverage ranged from a low of 5.4% at one facility to a high of 102.7% (note: 102.7% is the reported figure) at another facility.[c]
iii. Among 14 hospitals in New York City and Michigan, representing 36,046 live births during 2013, 77.3% of infants received the birth dose of Hepatitis B vaccine, when parent/guardian refusals were excluded from the denominator. This percentage represents an increase from 72.6% of infants during 2012.[d]
iv. The results of a Hepatitis B birth dose "Best Practices" Survey were presented at the National Immunization Conference in 2011. The survey was conducted by the New York State Department of Health´s Perinatal Hepatitis B Prevention Program to identify common practices among hospitals with the highest birth dose coverage. Best practices that increased or were associated with high birth dose adherence included: early parental education prior to hospitalization, early consent before or upon hospital admission, staff education and "buy in", and state-funded vaccine for the birth dose at no cost to hospitals (a universal Hepatitis B vaccine supply policy).[e]
v. A Public Health Evaluation Project at 119 hospitals in Texas consisting of a chart review during 2009-2010 demonstrated the 0-3 day of life Hepatitis B vaccination rates by birthing facility: mean 90.4%; median 95.5%; maximum 100%; minimum 21.2%.[f]
vi. A survey of a nationally representative sample of birthing hospitals was conducted in 2005, with review of over 10,000 mother/infant charts. The study described major gaps in hospital policies and practices designed to prevent perinatal transmission of Hepatitis B virus. Receipt of Hepatitis B vaccine within 12 hours of birth, as recommended by the Advisory Committee on Immunization Practices (ACIP), was confirmed in 67.1% of infants born to HBsAg-positive women. More than one-tenth (13.7%) of infants born to HBsAg-positive women (infants at highest risk of perinatal Hepatitis B virus transmission) received no Hepatitis B

vaccine prior to hospital discharge. Overall, 69% of infants born to HBsAg-negative pregnant women received the birth dose prior to hospital discharge. The existence of a written policy was the strongest correlate of adherence to birth dose administration of newborn infants.[g]

vii. During July 1999-October 2002, a survey of public health departments reported more than 500 hospital medical errors in failures to administer immunoprophylaxis at birth, including routine Hepatitis B birth dose was not part of hospital policy so that proper prophylaxis was not provided to infants.[h]

1b.3. **If no or limited performance data on the measure as specified is reported in 1b2, then provide a summary of data from the literature that indicates opportunity for improvement or overall less than optimal performance on the specific focus of measurement.**

a. Feasibility Study report based on review of medical records.
b. CDC. National, state, and selected local area vaccination coverage among children aged 19-35 months—United States, 2014. MMWR 2015; 64(33):889-96.
c. Hepatitis B Birth Dose Coverage by NYC Facility. http://www.nyc.gov/html/doh/html/hcp/cd-hepatitisb-pregnancy.shtml
d. Hepatitis B birth dose coverage data, excluding parent refusals, from New York City and Michigan hospitals; CDC, unpublished data, 2012-2013.
e. Pollock L. Hepatitis B Birth Dose Best Practices 2010 Survey. 2011 National Immunization Conference presentation (March 31, 2011) http://cdc.confex.com/cdc/nic2011/webprogram/Paper25179.html
f. Unpublished data from Texas Department of State Health Services, Public Health Evaluation Project (PHEP), 2010.
g. Willis BC, Wortley P, Wang S, Jacques-Carroll L, Zhang F. Gaps in Hospital Policies and Practices to Prevent Perinatal Transmission of Hepatitis B Virus. Pediatrics. 2010 125:704-711 (http://pediatrics.aappublications.org/content/125/4/704.full.html) alternate link: http://pediatrics.aappublications.org/content/125/4/704.long
h. http://www.immunize.org/catg.d/p2128.pdf and http://www.immunize.org/catg.d/p2062.pdf (Specific examples of medical errors resulting in HBV infection)

1b.4. **Provide disparities data from the measure as specified (current and over time) by population group, e.g., by race/ethnicity, gender, age, insurance status, socioeconomic status, and/or disability.** *(This is required for endorsement maintenance. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities include.) This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*

a. An analysis of data from 9,252 HBsAg-positive mothers and their infants among five U.S. jurisdictions during 2007-2013 demonstrated disparities regarding perinatal Hepatitis B transmission. Perinatal transmission occurred in 1.4% of infants born to mothers whose race was Asian/Pacific Islander, compared to 0.5%, 0.1%, and 0.6% of mothers whose race/ethnicity was black, white, or Hispanic, respectively ($p<0.01$). Perinatal transmission occurred in 1.3% of infants whose mothers were foreign-born, compared to 0.6% of infants whose mothers were U.S.-born ($p=0.02$).[a]
b. A 2006 study that estimated the number of births in the United States to HBsAg-positive women evaluated vital statistics data for 22 states that had information on country of birth of pregnant women. Results indicated that foreign-born women from countries highly endemic for Hepatitis B infection (despite being a minority of all women giving birth), U.S.- and Canadian-born non-Hispanic blacks, and Asian/Pacific Islanders represented the majority of all births to HBsAg-positive women. Of 2,359,912 births in the 22 states evaluated, approximately 16,500 births were estimated to be from HBV-infected women;80.6% of these were foreign-born women.[b]

1b.5. **If no or limited data on disparities from the measure as specified is reported in 1b4, then provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement. Include citations.**

a. Schillie S, Walker T, Veselsky S, Crowley S, Dusek C, Lazaroff J, Morris S, Onye K, Ko S, Fenlon N, Nelson N, Murphy T. Outcomes of Infants Born to Women Infected With Hepatitis B, Pediatrics 2015;135:e1141-7.
b. Din E, Wasley A, Jacques-Carroll L, Sirotkin B, Wang S. Estimating the Number of Births to Hepatitis B Virus-infected Women in 22 States, 2006;Pediatr Infect Dis J 2011;30:1-5.

1c. High Priority (previously referred to as High Impact)

The measure addresses:

- a specific national health goal/priority identified by DHHS or the National Priorities Partnership convened by NQF; OR
- a demonstrated high-priority (high-impact) aspect of healthcare (e.g., affects large numbers of patients and/or has a substantial impact for a smaller population; leading cause of morbidity/mortality; high resource use (current and/or future); severity of illness; and severity of patient/societal consequences of poor quality).

1c.1. **Demonstrated high priority aspect of healthcare**

Affects large numbers, A leading cause of morbidity/mortality, Frequently performed procedure, Patient/societal consequences of poor quality, Severity of illness

1c.2. If Other:

1c.3. Provide epidemiologic or resource use data that demonstrates the measure addresses a high priority aspect of healthcare. List citations in 1c.4.

a. Hepatitis B virus causes acute and chronic infections.

b. Perinatal Hepatitis B transmission rates vary with infectivity of the pregnant mother. Women with high infectivity (high viral loads) transmit the virus to approximately 90% of their infants; women with lower viral loads may transmit to 5-20% of their infants. Without intervention, transmission occurs in about 40% of pregnancies of women with chronic Hepatitis B infection.

c. Most morbidity and mortality associated with perinatal Hepatitis B infection occurs among infants who develop chronic Hepatitis B. Approximately 90% of infants with perinatal Hepatitis B will develop chronic infection, and about 25% of these infants will have premature death from complications of the chronic infection (i.e., cirrhosis, liver failure, and hepatocellular carcinoma). Without intervention, additional morbidity and mortality accrues from severe acute Hepatitis B among infants whose mothers have lower viral load, but still transmit the infection. These infants, although far less common, suffer fulminant acute Hepatitis B with very high mortality rates.

i. In 2008, an estimated 25,000 infants were born to Hepatitis B-infected mothers.

ii. Without vaccination (and Hepatitis B Immune Globulin [HBIG]), 6000-9000 of these infants would become chronically infected with Hepatitis B and approximately 2550 would be expected to die of chronic liver disease

iii. With delayed vaccination, additional uninfected infants will become infected after birth through exposure to household contacts with chronic Hepatitis B.

iv. Because the majority of chronically infected infants and children are asymptomatic, children with chronic infection will be identified only if tested for HBsAg. Routine HBsAg testing of infants and children is not done in the United States unless the infant is known to be born to an HBsAg-positive pregnant woman. Children and adults might be tested when they develop symptoms of liver disease.

d. The primary goal of Hepatitis B immunization starting at birth is to prevent chronic Hepatitis B infection when the risk is highest (birth through 5 years of age):

--When acute infection occurs at age less than 1 year, symptomatic infection occurs in less than 1% and chronic disease develops in 90%

--When acute infection occurs at age 1-5 years, symptomatic infection occurs in 5-15% and chronic disease develops in 25-50%

--When acute infection occurs at age greater than 5 years, symptomatic infection occurs in 20-50% and chronic disease develops in 6-10%

e. There are 2 common modes of Hepatitis B transmission during infancy and early childhood: 1) transmission from an infected mother to her infant, usually during delivery (perinatal/"vertical"), and 2) transmission from an infected (usually asymptomatic) household contact (“horizontal”). Both modes of transmission can be prevented by vaccination of infants starting at birth.

f. Vaccine efficacy in preventing perinatal Hepatitis B transmission:

i. Without post-exposure prophylaxis, approximately 90% of infants of HBsAg-positive mothers will develop chronic Hepatitis B infection.

ii. PEP includes Hepatitis B vaccine and HBIG, and is 85-95% effective in preventing perinatal Hepatitis B transmission when administered within 12-24 hours of birth, followed by completion of a Hepatitis B immunization series.[i] Current ACIP recommendations are for administration of Hepatitis B vaccine and HBIG within 12 hours of birth when the mother is known to be HBsAg-positive.

iii. Hepatitis B vaccine alone (without HBIG) starting at birth prevents transmission in 70-95% of infants.[ii-vii] Hepatitis B vaccine alone starting within the first 24 hours after birth is used in many countries to prevent perinatal Hepatitis B transmission.

iv. The major determinant of the effectiveness of post-exposure prophylaxis is early administration of the initial dose of vaccine. Studies are limited on the maximum interval after exposure during which post-exposure prophylaxis is effective, but it is unlikely to exceed 7 days.[viii-xi] The ACIP recommendation states the first dose of Hepatitis B vaccine should be given to all infants born to HBsAg-negative women before the infant is discharged from the hospital or birthing facility, which for most infants is before the 3rd day of life.

g. Prevention of Hepatitis B transmission from an infected household contact to an infant or child is a critical aspect of eliminating Hepatitis B transmission to vulnerable infants and children. Before perinatal Hepatitis B prevention programs, studies showed that 61-66% of children with chronic Hepatitis B infection were born to uninfected mothers and acquired the infection from close contacts, usually from members of the household.[xii, xiii] Hepatitis B vaccine starting at birth with completion of the Hepatitis B vaccine series will prevent these early childhood acquired infections.

h. Universal administration of a Hepatitis B vaccine “birth dose” is the safety net that prevents chronic Hepatitis B infection and life-long sequelae among infants born to HBsAg-positive pregnant women not identified due to lack of prenatal care, errors in testing, failure to report test results to public health, and failure to administer timely post-exposure prophylaxis, and for infants exposed to

Hepatitis B virus from infected household contacts before protection through routine Hepatitis B vaccination starting at a later age.
i. A recent analysis of infants born to infected mothers during 2007-2013 in five U.S. jurisdictions reaffirm the efficacy of prophylaxis in preventing perinatal transmission; among 9,252 infants, most of whom received prophylaxis, 1% developed perinatal Hepatitis B infection.[xiv]

1c.4. Citations for data demonstrating high priority provided in 1a.3

i. Andre FE, Zuckerman AJ. Review: protective efficacy of hepatitis B vaccines in neonates. J Med Virol 1994;44:144-51.
ii. Stevens CE, Neurath RA, Beasley RP, Szmuness W. HBeAg and anti-HBe detection by radioimmunoassay: correlation with vertical transmission of hepatitis B virus in Taiwan. J Med Virol 1979;3:237-41.
iii. Beasley RP, Hwang LY, Lee GC, et al. Prevention of perinatally transmitted hepatitis B virus infections with hepatitis B virus infections with hepatitis B immune globulin and hepatitis B vaccine. Lancet 1983;2(8359):1099-102.
iv. Lo KJ, Tsai YT, Lee SD, et al. Immunoprophylaxis of infection with hepatitis B virus in infants born to hepatitis B surface antigen-positive carrier mothers. J Infect Dis 1985;152:817-22.
v. Poovorawan Y, Sanpavat S, Pongpunlert W, et al. Comparison of a recombinant DNA hepatitis B vaccine alone or in combination with hepatitis B immune globulin for the prevention of perinatal acquisition of hepatitis B carriage. Vaccine 1990 (Suppl 8):S56-9.
vi. Assateerawatt A, Tanphaichitr VS, Suvatte V, In-ngarm L. Immunogenicity and protective efficacy of low dose recombinant DNA hepatitis B vaccine in normal and high-risk neonates. Asian Pac J Allergy Immunol 1991;9:89-93.
vii. Milne A, West DJ, Chinh DV, Moyes CD, Poerschke G. Field evaluation of the efficacy and immunogenicity of recombinant hepatitis B vaccine without HBIG in newborn Vietnamese infants. J Med Virol 2002;67:327-33.
viii. Krugman S, Overby LR, Mushahwar IK, Ling CM, Frosner GG, Deinhardt F. Viral hepatitis, type B: studies on natural history and prevention re-examined. N Engl J Med 1979;300:101-6.
ix. Grady GF. Viral hepatitis: passive prophylaxis with globulins---state of the art in 1978. In: Vyas GN, Cohen SN, Schmid R, eds. Viral hepatitis: a contemporary assessment of etiology, epidemiology, pathogenesis, and prevention. Philadelphia, PA: Franklin Institute Press, 1978:467-76.
x. Beasley RP, Stevens CE. Vertical transmission of HBV and interruption with globulin. In: Vyas GN, Cohen SN, Schmid R, eds. Viral hepatitis: a contemporary assessment of etiology, epidemiology, pathogenesis, and prevention. Philadelphia, PA: Franklin Institute Press; 1978:333-45.
xi. Marion SA, Tomm PM, Pi DW, Mathias RG. Long-term follow-up of hepatitis B vaccine in infants of carrier mothers. Am J Epidemiol 1994;140:734-46.
xii. Hurie MB, Mast EE, Davis JP. Horizontal transmission of hepatitis B virus infection to United States-born children of Hmong refugees. Pediatrics 1992;89:269-73.
xiii. Mahoney FJ, Lawrence M, Scott C, Le Q, Lambert S, Farley TA. Continuing risk for hepatitis B virus transmission among Southeast Asian infants in Louisiana. Pediatrics 1995;96:1113-6.
xiv. Schillie S, Walker T, Veselsky S, et al. Outcomes of Infants Born to Women Infected with Hepatitis B. Pediatrics 2015;135:e1141-1147.

1c.5. If a PRO-PM (e.g. HRQoL/functional status, symptom/burden, experience with care, health-related behaviors), provide evidence that the target population values the measured PRO and finds it meaningful. *(Describe how and from whom their input was obtained.)*

N/A

2. Reliability and Validity—Scientific Acceptability of Measure Properties

Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. ***Measures must be judged to meet the subcriteria for both reliability and validity to pass this criterion and be evaluated against the remaining criteria.***

2a.1. Specifications The measure is well defined and precisely specified so it can be implemented consistently within and across organizations and allows for comparability. eMeasures should be specified in the Health Quality Measures Format (HQMF) and the Quality Data Model (QDM).

De.5. **Subject/Topic Area** *(check all the areas that apply):*
Cancer : Liver, Infectious Diseases : Hepatitis, Infectious Diseases : Immunization, Perinatal and Reproductive Health, Perinatal and Reproductive Health : Newborn, Perinatal and Reproductive Health : Perinatal, Prevention, Prevention : Immunization

De.6. **Cross Cutting Areas** *(check all the areas that apply):*
Prevention, Prevention : Immunization

S.1. **Measure-specific Web Page** *(Provide a URL link to a web page specific for this measure that contains current detailed specifications including code lists, risk model details, and supplemental materials. Do not enter a URL linking to a home page or to general information.)*
http://www.cdc.gov/hepatitis/partners/perihepbcoord.htm

S.2a. **If this is an eMeasure**, HQMF specifications must be attached. Attach the zipped output from the eMeasure authoring tool (MAT) - if the MAT was not used, contact staff. (Use the specification fields in this online form for the plain-language description of the specifications)
This is not an eMeasure **Attachment:** HepBNewborn_Artifacts_09252013-635319495478389727-635627868869221191-635787044983961955.zip

S.2b. **Data Dictionary, Code Table, or Value Sets** *(and risk model codes and coefficients when applicable) must be attached. (Excel or csv file in the suggested format preferred - if not, contact staff)*
Attachment **Attachment:** ICD_9_and_10_code_tables-635320538568917230-635627868862357763-635787044982869948.pdf

S.3. **For endorsement maintenance**, please briefly describe any changes to the measure specifications since last endorsement date and explain the reasons.
No changes are being made to the measure specifications.

S.4. **Numerator Statement** *(Brief, narrative description of the measure focus or what is being measured about the target population, i.e., cases from the target population with the target process, condition, event, or outcome)*
IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
The number of live newborn infants administered Hepatitis B vaccine prior to discharge (or within 1 month of life, if the infant had an extended hospital stay)from the hospital/birthing facility ("birth dose" of Hepatitis B vaccine).

S.5. **Time Period for Data** *(What is the time period in which data will be aggregated for the measure, e.g., 12 mo, 3 years, look back to August for flu vaccination? Note if there are different time periods for the numerator and denominator.)*
One calendar year.

S.6. **Numerator Details** *(All information required to identify and calculate the cases from the target population with the target process, condition, event, or outcome such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
IF an OUTCOME MEASURE, describe how the observed outcome is identified/counted. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
Per hospital/birthing facility, the number of live newborn infants, during a calendar year, who received a dose of Hepatitis B vaccine prior to hospital/birthing facility discharge (or within 1 month of life, if the infant had an extended hospital stay). Acceptable data sources include: pharmacy records, vaccine consent forms, medication administration records, claims data, nurses notes, electronic medical records, or other available records.
a. Suggested ICD-9 code V05.3 converts to ICD-10 code z23 (type of immunization given will be identified by the procedure code—effective October 1, 2013. Procedure code for viral hepatitis unknown. Suggest the use of ICD-10 code z23.9955 described as

"prophylactic administration of vaccine against other diseases" or ICD-10 code z23.9959 described as "other vaccination or inoculation"): http://www.icd10data.com/ICD10CM/Codes/Z00-Z99/Z20-Z28/Z23-/Z23
b. CPT administration codes: 90744 (Hepatitis B vaccine) and 90471 (immunization administration code)

S.7. **Denominator Statement** *(Brief, narrative description of the target population being measured)*
The number of live newborn infants born at the hospital/birthing facility during the reporting window (one calendar year).

S.8. **Target Population Category** *(Check all the populations for which the measure is specified and tested if any):*
Children's Health, Populations at Risk

S.9. **Denominator Details** *(All information required to identify and calculate the target population/denominator such as definitions, specific data collection items/responses , code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
a. The number of live births at the hospital/birthing facility during one calendar year can be determined from a variety of sources, including the paper or electronic patient records, nursery birth records, or other available records. ICD-10 codes can be used. Stillborn deliveries are not included in the definition of the measure.
i. ICD-10 codes to be used (link: http://www.icd10data.com/ICD10CM/Codes/Z00-Z99/Z30-Z39/Z37-/#Z37 and http://www.icd10data.com/ICD10CM/Codes/Z00-Z99/Z30-Z39/Z38-/#Z38):

1. Z37.0 Single live birth
2. Z37.2 Twins, both live born
3. Z37.3 Twins, one live born and one stillborn
4. Z37.50 Multiple births, unspecified, all live born
5. Z37.51 Triplets, all live born
6. Z37.52 Quadruplets, all live born
7. Z37.53 Quintuplets, all live born
8. Z37.54 Sextuplets, all live born
9. Z37.59 Other multiple births, all live born
10. Z37.60 Multiple births, unspecified, some live born
11. Z37.61 Triplets, some live born
12. Z37.62 Quadruplets, some live born
13. Z37.63 Quintuplets, some live born
14. Z37.64 Sextuplets, some live born
15. Z37.69 Other multiple births, some live born
16. Z38.00 Single live born infant, delivered vaginally
17. Z38.01 Single live born infant, delivered by cesarean
18. Z38.1 Single live born infant, born outside hospital
19. Z38.2 Single live born infant, unspecified as to place of birth
20. Z38.30 Twin live born infant, delivered vaginally
21. Z38.31 Twin live born infant, delivered by cesarean
22. Z38.4 Twin live born infant, born outside hospital
23. Z38.5 Twin live born infant, unspecified as to place of birth
24. Z38.61 Triplet live born infant, delivered vaginally
25. Z38.62 Triplet live born infant, delivered by cesarean
26. Z38.63 Quadruplet live born infant, delivered vaginally
27. Z38.64 Quadruplet live born infant, delivered by cesarean
28. Z38.65 Quintuplet live born infant, delivered vaginally
29. Z38.66 Quintuplet live born infant, delivered by cesarean
30. Z38.68 Other multiple live born infant, delivered vaginally
31. Z38.69 Other multiple live born infant, delivered by cesarean
32. Z38.7 Other multiple live born infant, born outside hospital
33. Z38.8 Other multiple live born infant, unspecified as to place of birth

The results of this measure will identify that the coverage excludes infants whose parent/guardian refused Hepatitis B vaccine for their infant before hospital or facility discharge (or by 1 month of age if during a prolonged stay).

S.10. **Denominator Exclusions** *(Brief narrative description of exclusions from the target population)*

a. Determine number of live newborn infants born at the hospital/birthing facility whose parent/guardian refused Hepatitis B birth dose and exclude from the denominator. ICD-10 code for this information will include the following (link: http://www.icd10data.com/ICD10CM/Codes/Z00-Z99/Z20-Z28/Z28-/#Z28):
i. Z28.82 Immunization not carried out because of caregiver refusal

S.11. **Denominator Exclusion Details** *(All information required to identify and calculate exclusions from the denominator such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
Subtract from the number of infants discharged from the hospital/birthing facility, the number of infants born at the facility during one calendar year whose parent/guardian refused administration of a birth dose of Hepatitis B vaccine before discharge (or by 1 month of age if during a prolonged stay) from the hospital/birthing facility. Information on exclusions might come from a variety of sources, including vaccine consent forms, clinical notes, and medication administration records. No billing codes exist for vaccine refusal; therefore ICD-10 code Z28.82 should be used to document vaccine refusal.

S.12. **Stratification Details/Variables** *(All information required to stratify the measure results including the stratification variables, definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format with at S.2b)*
N/A

S.13. **Risk Adjustment Type** (Select type. Provide specifications for risk stratification in S.12 and for statistical model in S.14-15)
No risk adjustment or risk stratification
If other:

S.14. **Identify the statistical risk model method and variables** *(Name the statistical method - e.g., logistic regression and list all the risk factor variables. Note - risk model development and testing should be addressed with measure testing under Scientific Acceptability)*
N/A

S.15. **Detailed risk model specifications** *(must be in attached data dictionary/code list Excel or csv file. Also indicate if available at measure-specific URL identified in S.1.)*
Note: Risk model details (including coefficients, equations, codes with descriptors, definitions), should be provided on a separate worksheet in the suggested format in the Excel or csv file with data dictionary/code lists at S.2b.

S.15a. **Detailed risk model specifications** (*if not provided in excel or csv file at S.2b*)
N/A

S.16. **Type of score:**
Rate/proportion
If other:

S.17. **Interpretation of Score** *(Classifies interpretation of score according to whether better quality is associated with a higher score, a lower score, a score falling within a defined interval, or a passing score)*
Better quality = Higher score

S.18. **Calculation Algorithm/Measure Logic** *(Describe the calculation of the measure score as an ordered sequence of steps including identifying the target population; exclusions; cases meeting the target process, condition, event, or outcome; aggregating data; risk adjustment; etc.)*
a. Determine the number of live newborn infants at each hospital/birthing facility during one calendar year
b. Determine the number of live newborn infants born at the same hospital/birthing facility during the same calendar year who received a dose of Hepatitis B vaccine before hospital discharge (or by 1 month of age if during a prolonged stay)
c. Determine the number of parental/guardian refusals of Hepatitis B birth dose
d. Divide the number of live newborn infants born at the same hospital/birthing facility during the same time period who received a dose of Hepatitis B vaccine before hospital discharge (or by 1 month of age if during a prolonged stay)(b), by the number of live newborns at the same hospital/birthing facility during the same time period(a) minus those who were not vaccinated because of parent/guardian refusal of Hepatitis B birth dose(c)[b/(a-c)].

S.19. **Calculation Algorithm/Measure Logic Diagram URL or Attachment** *(You also may provide a diagram of the Calculation*

Algorithm/Measure Logic described above at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)
No diagram provided

S.20. **Sampling** *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*
IF a PRO-PM, identify whether (and how) proxy responses are allowed.
N/A, survey based on actual numbers.

S.21. **Survey/Patient-reported data** *(If measure is based on a survey, provide instructions for conducting the survey and guidance on minimum response rate.)*
IF a PRO-PM, specify calculation of response rates to be reported with performance measure results.
N/A

S.22. **Missing data** (specify how missing data are handled, e.g., imputation, delete case.)
Required for Composites and PRO-PMs.
Patients for whom records lack data indicating administration of a Hepatitis B vaccine birth dose will be considered to have not received a Hepatitis B vaccine birth dose. Patients for whom records lack data indicating parent/guardian refusal of a Hepatitis B vaccine birth dose will be considered to not have parent/guardian refusal. In other situations in which data is missing making it not possible to calculate the measure, that case will be deleted from the measure.

S.23. **Data Source** *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*
If other, please describe in S.24.
Electronic Clinical Data, Electronic Clinical Data : Electronic Health Record, Electronic Clinical Data : Pharmacy, Electronic Clinical Data : Registry, Other, Paper Medical Records

S.24. **Data Source or Collection Instrument** *(Identify the specific data source/data collection instrument e.g. name of database, clinical registry, collection instrument, etc.)*
IF a PRO-PM, identify the specific PROM(s); and standard methods, modes, and languages of administration.
N/A

S.25. **Data Source or Collection Instrument** *(available at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
No data collection instrument provided

S.26. **Level of Analysis** *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*
Facility

S.27. **Care Setting** *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*
Hospital/Acute Care Facility
If other:

S.28. **COMPOSITE Performance Measure** - Additional Specifications *(Use this section as needed for aggregation and weighting rules, or calculation of individual performance measures if not individually endorsed.)*
N/A

2a. Reliability – See attached Measure Testing Submission Form
2b. Validity – See attached Measure Testing Submission Form
testing_attachment-635930264661113192.docx

Measure Number (*if previously endorsed*)**:** 0475

Measure Title: Hepatitis B Vaccine Coverage Among All Live Newborn Infants Prior to Hospital or Birthing Facility Discharge

Date of Submission: 2/16/2016

Type of Measure:

☐ Composite – ***STOP – use composite testing form***	☐ Outcome (*including PRO-PM*)
☐ Cost/resource	☒ Process
☐ Efficiency	☐ Structure

Instructions

- Measures must be tested for all the data sources and levels of analyses that are specified. ***If there is more than one set of data specifications or more than one level of analysis, contact NQF staff*** about how to present all the testing information in one form.
- **For all measures, sections 1, 2a2, 2b2, 2b3, and 2b5 must be completed.**
- **For outcome and resource use measures**, section **2b4** also must be completed.
- If specified for **multiple data sources/sets of specificaitons** (e.g., claims and EHRs), section **2b6** also must be completed.
- Respond to all questions as instructed with answers immediately following the question. All information on testing to demonstrate meeting the subcriteria for reliability (2a2) and validity (2b2-2b6) must be in this form. An appendix for *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 20 pages (*incuding questions/instructions;* minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.
- For information on the most updated guidance on how to address sociodemographic variables and testing in this form refer to the release notes for version 6.6 of the Measure Testing Attachment.

Note: The information provided in this form is intended to aid the Steering Committee and other stakeholders in understanding to what degree the testing results for this measure meet NQF's evaluation criteria for testing.

2a2. Reliability testing [10] demonstrates the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise. For **PRO-PMs and composite performance measures**, reliability should be demonstrated for the computed performance score.

2b2. Validity testing [11] demonstrates that the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality. For **PRO-PMs and composite performance measures**, validity should be demonstrated for the computed performance score.

2b3. Exclusions are supported by the clinical evidence; otherwise, they are supported by evidence of sufficient frequency of occurrence so that results are distorted without the exclusion; [12]
AND
If patient preference (e.g., informed decision making) is a basis for exclusion, there must be evidence that the exclusion impacts performance on the measure; in such cases, the measure must be specified so that the information about patient preference and the effect on the measure is transparent (e.g., numerator category computed separately,

denominator exclusion category computed separately). [13]

2b4. For outcome measures and other measures when indicated (e.g., resource use):

- **an evidence-based risk-adjustment strategy** (e.g., risk models, risk stratification) is specified; is based on patient factors (including clinical and sociodemographic factors) that influence the measured outcome and are present at start of care; [14,15] and has demonstrated adequate discrimination and calibration

OR

- rationale/data support no risk adjustment/ stratification.

2b5. Data analysis of computed measure scores demonstrates that methods for scoring and analysis of the specified measure allow for **identification of statistically significant and practically/clinically meaningful [16] differences in performance**;

OR

there is evidence of overall less-than-optimal performance.

2b6. If multiple data sources/methods are specified, there is demonstration they produce comparable results.

2b7. For **eMeasures, composites, and PRO-PMs** (or other measures susceptible to missing data), analyses identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias.

Notes

10. Reliability testing applies to both the data elements and computed measure score. Examples of reliability testing for data elements include, but are not limited to: inter-rater/abstractor or intra-rater/abstractor studies; internal consistency for multi-item scales; test-retest for survey items. Reliability testing of the measure score addresses precision of measurement (e.g., signal-to-noise).

11. Validity testing applies to both the data elements and computed measure score. Validity testing of data elements typically analyzes agreement with another authoritative source of the same information. Examples of validity testing of the measure score include, but are not limited to: testing hypotheses that the measures scores indicate quality of care, e.g., measure scores are different for groups known to have differences in quality assessed by another valid quality measure or method; correlation of measure scores with another valid indicator of quality for the specific topic; or relationship to conceptually related measures (e.g., scores on process measures to scores on outcome measures). Face validity of the measure score as a quality indicator may be adequate if accomplished through a systematic and transparent process, by identified experts, and explicitly addresses whether performance scores resulting from the measure as specified can be used to distinguish good from poor quality.

12. Examples of evidence that an exclusion distorts measure results include, but are not limited to: frequency of occurrence, variability of exclusions across providers, and sensitivity analyses with and without the exclusion.

13. Patient preference is not a clinical exception to eligibility and can be influenced by provider interventions.

14. Risk factors that influence outcomes should not be specified as exclusions

15. With large enough sample sizes, small differences that are statistically significant may or may not be practically or clinically meaningful. The substantive question may be, for example, whether a statistically significant difference of one percentage point in the percentage of patients who received smoking cessation counseling (e.g., 74 percent v. 75 percent) is clinically meaningful; or whether a statistically significant difference of $25 in cost for an episode of care (e.g., $5,000 v. $5,025) is practically meaningful. Measures with overall less-than-optimal performance may not demonstrate much variability across providers.

1. DATA/SAMPLE USED FOR ALL TESTING OF THIS MEASURE

Often the same data are used for all aspects of measure testing. In an effort to eliminate duplication, the first five questions apply to all measure testing. If there are differences by aspect of testing,(e.g., reliability vs. validity) be sure to indicate the specific differences in question 1.7.

1.1. What type of data was used for testing? (*Check all the sources of data identified in the measure specifications and data used for testing the measure. Testing must be provided for all the sources of data specified and intended for*

measure implementation. ***If different data sources are used for the numerator and denominator, indicate N [numerator] or D [denominator] after the checkbox.***)

Measure Specified to Use Data From: (*must be consistent with data sources entered in S.23*)	**Measure Tested with Data From:**
☒ abstracted from paper record	☒ abstracted from paper record
☐ administrative claims	☐ administrative claims
☒ clinical database/registry	☒ clinical database/registry
☒ abstracted from electronic health record	☒ abstracted from electronic health record
☐ eMeasure (HQMF) implemented in EHRs	☐ eMeasure (HQMF) implemented in EHRs
☒ other: consent form, delivery/nursery log, reporting mom to health department, newborn nursery, Quality Dashboard, medical dispensing logs, state electronic birth certificate program	☒ other: consent form, delivery/nursery log, reporting mom to health department, newborn nursery, Quality Dashboard, medical dispensing logs, state electronic birth certificate program

1.2. If an existing dataset was used, identify the specific dataset (*the dataset used for testing must be consistent with the measure specifications for target population and healthcare entities being measured; e.g., Medicare Part A claims, Medicaid claims, other commercial insurance, nursing home MDS, home health OASIS, clinical registry*).

N/A

1.3. What are the dates of the data used in testing? Click here to enter date range

Reliability testing: January 1, 2010 – December 31, 2013
Validity testing: October 13, 2015 – October 23, 2015

1.4. What levels of analysis were tested? (*testing must be provided for <u>all</u> the levels specified and intended for measure implementation, e.g., individual clinician, hospital, health plan*)

Measure Specified to Measure Performance of: (*must be consistent with levels entered in item S.26*)	**Measure Tested at Level of:**
☐ individual clinician	☒ individual clinician (Validity testing [face validity])
☐ group/practice	☐ group/practice
☒ hospital/facility/agency	☒ hospital/facility/agency (Reliability testing)
☐ health plan	☐ health plan
☐ other: Click here to describe	☐ other: Click here to describe

1.5. How many and which <u>measured entities</u> were included in the testing and analysis (by level of analysis and data source)? (*identify the number and descriptive characteristics of measured entities included in the analysis (e.g., size, location, type); if a sample was used, describe how entities were selected for inclusion in the sample*)

The Feasibility Study was conducted in 50 hospitals that were representative of labor and delivery hospitals in Texas. The annual birth cohort size at Study hospitals ranged from 219 to 6,530 and together represented more than 100,000 births. Hospitals were divided among urban and rural locations and included for-profit, not-for-profit and public facilities. Thirty-six of 50 hospitals representing more than 62,000 births calculated the measure.

Reliability testing: Sixteen distinct hospitals and/or birthing facilities located in New York City and Michigan (respectively, 11 and 5 facilities) are included in the NQF 0475 measurement reliability assessment. Most facilities (87.5%, 14/16) are classified as non-government community hospitals and the remainder (12.5%, 2/16) are classified as a state/local government community hospital. Seventy-five percent of facilities are teaching hospitals. The level of neonatal care each facility provides varies: 18.75% are Level 1 with a well newborn nursery, 25.00% are Level II with a special care nursery, 37.50% are Level III with a Neonatal Intensive Care Unit (NICU), and 18.75% are Level IV with a regional NICU. There are different methods used by the facilities to report hepatitis B vaccine birth dose: 43.75%

reported on the newborn's electronic birth certificate and 56.25% reported electronically to an immunization registry (i.e. NYC DOHMH Citywide Immunization Registry for NYC or the MDCH Michigan Care Improvement Registry for Michigan). One facility located in NYC reports birth dose by both newborn electronic birth certificate and electronically to an immunization registry. The majority of facilities (81.25%) are located in an urban setting, 12.50% in a suburban setting, and 6.25% in a rural setting. Facility bed size in 2015 ranged from 62 to 843, with a mean of 347.75 and median of 273.50. Facility live births in 2014 are available for 11 facilities with a range of 183 to 8561, mean of 3243.45, and median of 2358.00; facility live births in 2011 are available for 5 facilities with a range of 314 to 1046, mean of 734.20, and median of 819.00. The proportion of patients admitted with Medicaid to facilities ranges from 12% to 67%, with a mean of 35.86%, and a median of 31.50%. The proportion of deliveries at facilities covered by Medicaid or other public-funded health insurance programs ranges from 11.00% to 98.00%, with a mean of 60.73%, and a median of 67.00%.

Validity testing: 14 medical, healthcare, and public health professionals with expertise in Hepatitis B responded to the validity survey; the response rate was 63.6% (14 of 22 persons comprising the Advisory Committee on Immunization Practices Hepatitis Work Group responded).

1.6. How many and which <u>patients</u> were included in the testing and analysis (by level of analysis and data source)? (*identify the number and descriptive characteristics of patients included in the analysis (e.g., age, sex, race, diagnosis); if a sample was used, describe how patients were selected for inclusion in the sample*)

Reliability testing: A convenience sample of hospitals in New York City and Michigan was chosen for the analysis. All live births from each hospital were included in the analysis; therefore, samples of patients (births) were not used from the hospitals to conduct the testing.

Validity testing: The survey was administered to the 9 liaison representatives and 13 federal employees comprising the Advisory Committee on Immunization Practices Hepatitis Work Group. The response rate was 63.6% (14/22).

1.7. If there are differences in the data or sample used for different aspects of testing (e.g., reliability, validity, exclusions, risk adjustment), identify how the data or sample are different for each aspect of testing reported below.

Reliability testing: A convenience sample of hospitals in New York City and Michigan was chosen for the analysis. All live births from each hospital were included in the analysis; therefore, samples of patients (births) were not used from the hospitals to conduct the testing.

Validity testing: The survey was administered to the 9 liaison representatives and 13 federal employees comprising the Advisory Committee on Immunization Practices Hepatitis Work Group. The response rate was 63.6% (14/22).

1.8 What were the patient-level sociodemographic (SDS) variables that were available and analyzed in the data or sample used? For example, patient-reported data (e.g., income, education, language), proxy variables when SDS data are not collected from each patient (e.g. census tract), or patient community characteristics (e.g. percent vacant housing, crime rate).

N/A

2a2. RELIABILITY TESTING

***<u>Note</u>**: If accuracy/correctness (validity) of data elements was empirically tested, separate reliability testing of data elements is not required – in 2a2.1 check critical data elements; in 2a2.2 enter "see section 2b2 for validity testing of data elements"; and skip 2a2.3 and 2a2.4.*

2a2.1. What level of reliability testing was conducted? (*may be one or both levels*)

☐ **Critical data elements used in the measure** (*e.g., inter-abstractor reliability; data element reliability must address ALL critical data elements*)

☒ **Performance measure score** (e.g., *signal-to-noise analysis*)

2a2.2. For each level checked above, describe the method of reliability testing and what it tests (*describe the steps—do not just name a method; what type of error does it test; what statistical analysis was used*)

Reliability was evaluated by the level of concordance between the rates calculated by the facilities in the Feasibility Study when compared with the rates determined by medical chart review in the PHEP.

Reliability was calculated as based on the methods outlined by John L. Adams, "The Reliability of Provider Profiling: A Tutorial" in the Rand technical report, TR-653-NCQA, 2009: http://www.rand.org/pubs/technical_reports/TR653.html.

1. The number of deliveries and the number of infants who received a birth dose at each hospital in the dataset was collected.
2. The birth dose coverage rate for each hospital was calculated and the number of infants whose parents or guardians refused the birth dose was subtracted from the denominator.
3. The SAS macro program (in RAND paper page 26, http://www.qistats.co.uk/BetaBinomial.html) was downloaded and executed to get the alpha and beta for the set of hospitals.
4. The between hospital ("provider") variance (sigma, squared) was calculated using alpha and beta derived from the RAND formula:

$$\sigma^2_{provider-to-provider} = \frac{\alpha\beta}{(\alpha+\beta+1)(\alpha+\beta)^2} \ .$$

5. The reliability was calculated for each hospital using the formula below from RAND, where *p*=birth dose rate and n=number of births.

$$reliability = \frac{\sigma^2_{provider-to-provider}}{\sigma^2_{provider-to-provider} + \sigma^2_{binomial}} = \frac{\sigma^2_{provider-to-provider}}{\sigma^2_{provider-to-provider} + \frac{p(1-p)}{n}} \ .$$

Reliability is the ratio of signal to noise; the signal is the proportion of measured variability that can be explained by real differences in performance. According to the Rand tutorial, this two-level hierarchical model separates the observed variability in hospital ("provider") scores into two components: variability between hospitals and variability within hospitals. The hierarchical model estimates the two variance components of reliability. The hospital-to-hospital variance is the variance of the true values, and the hospital-specific variance is due to measurement error. The hierarchical model conceptually removes measurement error variance from the true, hospital-to-hospital variance.

2a2.3. For each level of testing checked above, what were the statistical results from reliability testing? (e.*g., percent agreement and kappa for the critical data elements; distribution of reliability statistics from a signal-to-noise analysis*)

The baseline birth dose MEASURE as calculated by individual hospitals ranged from 8% - 100%. Analysis of these data compared with estimates from the chart reviews showed a variance of ± 10% for most hospitals that provided a result (30 of 36 hospitals). The difference between the hospital calculation of the MEASURE and the chart review estimate was substantially larger in the remaining 6 hospitals.

Distribution of reliability statistics from signal-to-noise analysis:
2013: mean=0.99844, median=0.99916, min=0.99251, max=0.99999 (14 hospitals; 32,471 live births)

2012: mean=0.99900, median=0.99944, min=0.99631, max=1.00000 (13 hospitals; 33,078 live births)
2011: mean=0.99834, median=0.99900, min=0.99660, max=0.99999 (11 hospitals; 24,914 live births)
2010: mean=0.99368, median=0.99388, min=0.98132, max=1.00000 (8 hospitals; 16,468 live births)

2a2.4 What is your interpretation of the results in terms of demonstrating reliability? (i.*e., what do the results mean and what are the norms for the test conducted?*)

Reliability is the ratio of signal to noise; the signal is the proportion of measured variability that can be explained by real differences in performance. Reliability can range from zero to one; a reliability of zero indicates variability is attributable to measurement error, while a reliability of one implies that variability is attributable to actual performance differences. According to the Rand tutorial, psychometricians consider a reliability of 0.90 an acceptable cut-off for making conclusions about individuals, while lower levels (0.70-0.80) are acceptable for making conclusions about groups. As such, measures from our analysis ranging from 0.98132 to 1.00000 indicate very high reliability, indicating that variability between hospitals regarding the Hepatitis B vaccine birth dose is due to actual performance differences rather than measurement error. Results support NQF endorsement of the measure.

_____

2b2. VALIDITY TESTING

2b2.1. What level of validity testing was conducted? (*may be one or both levels*)

☐ **Critical data elements** (*data element validity must address ALL critical data elements*)

☐ **Performance measure score**

☐ **Empirical validity testing**

☒ **Systematic assessment of face validity of <u>performance measure score</u> as an indicator** of quality or resource use (*i.e., is an accurate reflection of performance on quality or resource use and can distinguish good from poor performance*)

2b2.2. For each level of testing checked above, describe the method of validity testing and what it tests (*describe the steps—do not just name a method; what was tested, e.g., accuracy of data elements compared to authoritative source, relationship to another measure as expected; what statistical analysis was used)*

Validity of hospital-determined birth dose vaccination rate (the MEASURE) was determined by comparing the birth dose MEASURE result determined at each hospital with the birth dose rate estimated from review of a sample of infant birth charts at each hospital covering the same birth cohort. The minimum number of charts reviewed was determined from a table of sample sizes based on the expected hepatitis B birth dose coverage (range 50%-95%) and the hospital-specific size of the annual birth cohort (n = 100 - >20,000). The 2008 birth cohort size at each hospital was determined in a policy and practices survey in Texas. The estimated birth dose coverage rate was assumed to be 75% for all hospitals based on the Texas statewide coverage from the (2006) National Immunization Survey, and survey data from Dallas County, Texas. Charts were selected for review at each hospital using a sampling interval by date of birth depending on the total number of births annually (e.g., 1 chart for every 20 births). The number of infant charts reviewed per hospital ranged from 96-116 (average 106 records).

Validity testing (face validity) was tested in accordance with the methods outlined by Spertus et al., "American College of Cardiology and American Heart Association Methodology for the Selection and Creation of Performance Measures for Quantifying the Quality of Cardiovascular Care" in Circulation, 2005;111:1703-1712. Face validity was assessed systematically, in a transparent process, by recognized experts in the field to assess the computed performance measure score as a method to distinguish between good and poor quality.

1. A rating form (survey) assessing the measure using a 5 point Likert scale in terms of usefulness in improving patient outcomes, measure design, measure implementation, and overall assessment was developed, revised, and finalized.

2. Experts consisting of professionals from medicine, healthcare, and public health were identified for survey participation. Experts consisted of Advisory Committee on Immunization Practices Hepatitis Work Group members
3. The survey was administered electronically to 22 experts.
4. 14 persons responded to the survey (response rate=63.6%)
5. Descriptive results from completed surveys were tabulated, analyzed, and summarized.

2b2.3. What were the statistical results from validity testing? (*e.g., correlation; t-test*)

The baseline birth dose MEASURE as calculated by individual hospitals ranged from 8% - 100%. Analysis of these data compared with estimates from the chart reviews showed a variance of ± 10% for most hospitals that provided a result (30 of 36 hospitals). The difference between the hospital calculation of the MEASURE and the chart review estimate was substantially larger in the remaining 6 hospitals.

a. The birth dose of hepatitis B vaccine contributes to quality of care by providing a safety-net for infants who would not receive post-exposure prophylaxis because their mother´s chronic hepatitis B infection is not determined or detected, is misinterpreted or incorrectly recorded, or who return to a household with risk of transmission from family members with chronic hepatitis B infection (often unknown). Infants have a 90% chance of chronic hepatitis B if infected. The first dose of hepatitis B vaccine provides the initial step for prevention of almost certain life-long chronic hepatitis B infection with ~25% risk of cirrhosis, liver failure, and liver cancer. This is the critical "window" for prevention since chronic hepatitis B infection is not "curable".

Survey results reflect number of respondents indicating agreement

Numerator: The number of live newborn infants administered hepatitis B vaccine prior to discharge from the hospital/birthing facility ("birth dose" of hepatitis B vaccine).

Denominator: The number of live newborn infants born at the hospital/birthing facility during the reporting window (one calendar year), excluding infants whose parent(s)/guardian(s) refused hepatitis B vaccine.

	Strongly Disagree (1)	Disagree (2)	Neutral (3)	Agree (4)	Strongly Agree (5)	Mean Response
Useful in Improving Patient Outcomes						
(1) **Evidence-based**: The scientific basis of the measure is well established.	0	0	1	6	7	4.43
(2) **Interpretable**: The results of the measure are interpretable by practitioners.	0	1	0	8	5	4.21
(3) **Actionable**: The measure addresses an area that is under the practitioner's control.	0	1	1	5	7	4.29
Measure Design						
(2) **Denominator:** The patient group to whom this measure applies (denominator) is clinically	0	1	0	8	5	4.21

meaningful.						
(2) **Numerator**: The definition of the numerator for this measure is clinically meaningful.	0	1	0	6	7	4.36
(3) **Validity**: The measure appears to measure what it is intended to (face validity).	0	0	0	7	7	4.50
a. The measure captures most meaningful aspects of care (content validity).	0	0	2	6	6	4.29
b. The measure correlates well with other measures of the same aspect of care (construct validity).	0	0	3	5	6	4.21
Measure Implementation						
(1) **Feasibility**: The data required for the measure is likely to be obtained with reasonable effort.	0	1	1	7	5	4.14
a. The data required for the measure is likely to be obtained at reasonable cost.	0	0	3	7	4	4.07
b. The data required for the measure is likely to be obtained within the period allowed for data collection.	0	0	1	9	4	4.21
Overall Assessment						
	Must NOT Include (1)	**Should NOT Include (2)**	**Neutral (3)**	**Should Include (4)**	**Must Include (5)**	**Mean Response**
Considering your assessment of this measure on all dimensions above, rate this measure overall for inclusion into the NQF Measure #0475 "Hepatitis B Vaccine Coverage Among All Live Newborn Infants Prior to Hospital or Birthing Facility	0	1	1	5	7	4.29

Discharge".						

2b2.4. What is your interpretation of the results in terms of demonstrating validity? (i.*e., what do the results mean and what are the norms for the test conducted?*)

The domain-specific mean responses ranged from 4.07 to 4.50, corresponding to a mean response between "Agree" and "Strongly Agree." The domain assessing whether or not the measure appears to measure what it is intended to measure, based on the respondents' judgement of the clarity and comprehensiveness of the measure, had a mean response of 4.5, with half of respondents (7 of 14) strongly agreeing that the measure appears to measure what it is intended to, and the other half of respondents (7 of 14) agreeing that the measure appears to measure what it is intended to measure. Results from the validity survey support that the construct underlying the measure is associated with a meaningful process. Similarly, half (7 of 14) respondents indicated this measure must be included into the NQF process, while 35.7% (5 of 14) indicated it should be included. Face validity was assessed in a systematic and transparent process, as assessed by experts from the field. Results are consistent with high face validity and support NQF endorsement of the measure.

_____

2b3. EXCLUSIONS ANALYSIS

NA ☐ no exclusions — *skip to section 2b4*

2b3.1. Describe the method of testing exclusions and what it tests (*describe the steps—do not just name a method; what was tested, e.g., whether exclusions affect overall performance scores; what statistical analysis was used*)

A feasibility study (Feasibility Study) was conducted to evaluate adjusting the denominator to exclude infants whose parent/guardian refused the hepatitis B birth dose vaccination. ICD-9 codes (http://icd9cm.chrisendres.com/index.php?action=child&recordid=11296) used to determine parental/guardian refusal were V64.05—vaccination not carried out because of care giver/parental refusal (recorded in infant's medical chart) and/or V64.06—vaccination not carried out because of patient refusal (recorded in mother's medical chart). Thirty-eight percent of hospitals queried (representing >30,000 births) were able to make the adjustment for vaccine refusal at the time of the study (2010). The adjustment improved vaccination rates for some hospitals. About 50% of hospitals queried indicated they expected future improvements in electronic records that would facilitate or allow making this adjustment. However, currently, the adjustment relies on a variety of information sources and is not collected in a standardized manner. Some hospitals do not have the capacity to make this adjustment.

Comparison of birth dose rate (MEASURE) including refusals and excluding refusals

1. The number of deliveries and the number of infants who received a birth dose at each hospital in the dataset was collected.
2. The birth dose coverage rate for each hospital was calculated. The number of infants whose parents or guardians refused the birth dose was NOT subtracted from the denominator.
3. The SAS macro program (in RAND paper page 26, http://www.qistats.co.uk/BetaBinomial.html) was downloaded and executed to get the alpha and beta for the set of hospitals.
4. The between hospital ("provider") variance (sigma, squared) was calculated using alpha and beta derived from the RAND formula:

$$\sigma^2_{provider-to-provider} = \frac{\alpha\beta}{(\alpha+\beta+1)(\alpha+\beta)^2} \cdot$$

6. The reliability was calculated for each hospital using the formula below from RAND, where *p*=birth dose rate and n=number of births.

$$reliability = \frac{\sigma^2_{provider-to-provider}}{\sigma^2_{provider-to-provider} + \sigma^2_{binomial}} = \frac{\sigma^2_{provider-to-provider}}{\sigma^2_{provider-to-provider} + \frac{p(1-p)}{n}} .$$

Reliability is the ratio of signal to noise; the signal is the proportion of measured variability that can be explained by real differences in performance. According to the Rand tutorial, this two-level hierarchical model separates the observed variability in hospital ("provider") scores into two components: variability between hospitals and variability within hospitals. The hierarchical model estimates the two variance components of reliability. The hospital-to-hospital variance is the variance of the true values, and the hospital-specific variance is due to measurement error. The hierarchical model conceptually removes measurement error variance from the true, hospital-to-hospital variance.

2b3.2. What were the statistical results from testing exclusions? (*include overall number and percentage of individuals excluded, frequency distribution of exclusions across measured entities, and impact on performance measure scores*)

The difference in birth dose coverage result was compared between the MEASURE including refusals and the MEASURE without refusals. In the Feasibility Study, 16 hospitals calculated this difference in birth dose rate: mean ± 4.0%; median +1%; range -8% to + 25%. In the Texas Public Health Evaluation Project (PHEP) for the same 16 hospitals based on chart reviews the differences were: mean ±1.0%; median 0%; range -3% to + 7%. Refusal rates for birth dose of hepatitis B vaccine was estimated in Texas 2009-2010: mean 2.9%; median 1.0%; minimum 0.0%; maximum 21.6%.

Distribution of reliability statistics from signal-to-noise analysis, including refusals in the denominator:
2013: mean=0.99738, median=0.99865, min=0.98977, max=0.99999 (14 hospitals; 36,046 live births)
2012: mean=0.99852, median=0.99904, min=0.99446, max=1.00000 (13 hospitals; 36,790 live births)
2011: mean=0.99923, median=0.99957, min=0.99712, max=0.99999 (11 hospitals; 32,862 live births)
2010: mean=0.99856, median=0.99903, min=0.99525, max=0.99986 (8 hospitals; 23,275 live births)

2b3.3. What is your interpretation of the results in terms of demonstrating that exclusions are needed to prevent unfair distortion of performance results? (*i.e., the value outweighs the burden of increased data collection and analysis. Note: **If patient preference is an exclusion**, the measure must be specified so that the effect on the performance score is transparent, e.g., scores with and without exclusion*)

Reliability is the ratio of signal to noise; the signal is the proportion of measured variability that can be explained by real differences in performance. Reliability can range from zero to one; a reliability of zero indicates variability is attributable to measurement error, while a reliability of one implies that variability is attributable to actual performance differences. According to the Rand tutorial, psychometricians consider a reliability of 0.90 an acceptable cut-off for making conclusions about individuals, while lower levels (0.70-0.80) are acceptable for making conclusions about groups. As such, measures from our analysis, not excluding refusals from the denominator, ranging from 0.98977 to 1.00000 indicate very high reliability, indicating that variability between hospitals regarding the Hepatitis B vaccine birth dose is due to actual performance differences rather than measurement error. Results support consideration of removing exclusions from the denominator.

2b4. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES
If not an intermediate or health outcome, or PRO-PM, or resource use measure, skip to section 2b5.

2b4.1. What method of controlling for differences in case mix is used?
☒ **No risk adjustment or stratification**
☐ **Statistical risk model with** Click here to enter number of factors **risk factors**

☐ **Stratification by** Click here to enter number of categories **risk categories**
☐ **Other,** Click here to enter description

2b4.2. If an outcome or resource use measure is not risk adjusted or stratified, provide rationale and analyses to demonstrate that controlling for differences in patient characteristics (case mix) is not needed to achieve fair comparisons across measured entities.

The Hepatitis B vaccine birth dose is recommended universally, for all infants. The validity survey indicated that half (7 of 14) respondents strongly agreed that this measure addresses an area under the practitioner's control, and 35.7% (5 of 14) agreed it addresses in area under the practitioner's control. Hence, risk adjustment or stratification is not warranted for this measure, as the process is universally recommended for all infants.

2b4.3. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors (clinical factors or sociodemographic factors) used in the statistical risk model or for stratification by risk (*e.g., potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of p<0.10; correlation of x or higher; patient factors should be present at the start of care*)

N/A

2b4.4a. What were the statistical results of the analyses used to select risk factors?

N/A

2b4.4b. Describe the analyses and interpretation resulting in the decision to select SDS factors (e.g. prevalence of the factor across measured entities, empirical association with the outcome, contribution of unique variation in the outcome, assessment of between-unit effects and within-unit effects)

N/A

2b4.5. Describe the method of testing/analysis used to develop and validate the adequacy of the statistical model or stratification approach (*describe the steps—do not just name a method; what statistical analysis was used*)

Provide the statistical results from testing the approach to controlling for differences in patient characteristics (case mix) below.
If stratified, skip to 2b4.9

N/A

2b4.6. Statistical Risk Model Discrimination Statistics (*e.g., c-statistic, R-squared*)**:**

N/A

2b4.7. Statistical Risk Model Calibration Statistics (*e.g., Hosmer-Lemeshow statistic*):

N/A

2b4.8. Statistical Risk Model Calibration – Risk decile plots or calibration curves:

N/A

2b4.9. Results of Risk Stratification Analysis:

N/A

2b4.10. What is your interpretation of the results in terms of demonstrating adequacy of controlling for differences in patient characteristics (case mix)? (i.*e., what do the results mean and what are the norms for the test conducted*)

N/A

2b4.11. Optional Additional Testing for Risk Adjustment (*not required, but would provide additional support of adequacy of risk model, e.g., testing of risk model in another data set; sensitivity analysis for missing data; other methods that were assessed*)

N/A

_____

2b5. IDENTIFICATION OF STATISTICALLY SIGNIFICANT & MEANINGFUL DIFFERENCES IN PERFORMANCE

2b5.1. Describe the method for determining if statistically significant and clinically/practically meaningful differences in performance measure scores among the measured entities can be identified (*describe the steps—do not just name a method; what statistical analysis was used? Do not just repeat the information provided related to performance gap in 1b)*

Hepatitis B vaccine birth dose coverage, excluding (and not excluding) parent refusals from the denominator, was calculated for hospitals in New York City and Michigan. Descriptive statistics were used to compare coverage between facilities.

1. The number of deliveries and the number of infants who received a birth dose at each hospital in the dataset was collected.
2. The birth dose coverage rate for each hospital was calculated, excluding and not excluding the number of infants whose parents or guardians refused the birth dose from the denominator.

2b5.2. What were the statistical results from testing the ability to identify statistically significant and/or clinically/practically meaningful differences in performance measure scores across measured entities? (e.g., *number and percentage of entities with scores that were statistically significantly different from mean or some benchmark, different from expected; how was meaningful difference defined*)

Excluding refusals from denominator:

2013
Number of facilities: 14
Total number of live births, excluding refusals: 32,471
Minimum birth dose coverage: 29.95%
Maximum birth dose coverage: 99.96%
Mean birth dose coverage: 91.09%
Median birth dose coverage: 96.08%

2012
Number of facilities: 13
Total number of live births, excluding refusals: 33,078
Minimum birth dose coverage: 30.89%
Maximum birth dose coverage: 100.00%
Mean birth dose coverage: 87.38%
Median birth dose coverage: 95.47%

2011

Number of facilities: 11
Total number of live births, excluding refusals: 24,914
Minimum birth dose coverage: 32.37%
Maximum birth dose coverage: 99.91%
Mean birth dose coverage: 87.41%
Median birth dose coverage: 96.25%

2010
Number of facilities: 8
Total number of live births, excluding refusals: 16,468
Minimum birth dose coverage: 81.55%
Maximum birth dose coverage: 100.00%
Mean birth dose coverage: 91.73%
Median birth dose coverage: 94.19%

NOT excluding refusals from denominator:

2013
Number of facilities: 14
Total number of live births, not excluding refusals: 36,046
Minimum birth dose coverage: 29.95%
Maximum birth dose coverage: 99.96%
Mean birth dose coverage: 84.15%
Median birth dose coverage: 90.78%

2012
Number of facilities: 13
Total number of live births, not excluding refusals: 36,790
Minimum birth dose coverage: 30.89%
Maximum birth dose coverage: 100.00%
Mean birth dose coverage: 80.62%
Median birth dose coverage: 90.23%

2011
Number of facilities: 11
Total number of live births, not excluding refusals: 32,862
Minimum birth dose coverage: 25.61%
Maximum birth dose coverage: 99.91%
Mean birth dose coverage: 72.32%
Median birth dose coverage: 93.95%

2010
Number of facilities: 8
Total number of live births, not excluding refusals: 23,275
Minimum birth dose coverage: 33.88%
Maximum birth dose coverage: 98.67%
Mean birth dose coverage: 73.89%
Median birth dose coverage: 90.51%

2b5.3. What is your interpretation of the results in terms of demonstrating the ability to identify statistically significant and/or clinically/practically meaningful differences in performance across measured entities? (i.*e., what do*

the results mean in terms of statistical and meaningful differences?)

There is significant variability between hospitals for Hepatitis B vaccine birth dose coverage, ranging from a low of 29.95% to a high of 99.96% in 2013 among 14 facilities representing over 32,000 births excluding parent refusals (over 36,000 not excluding parent refusals). Furthermore, the mean birth dose coverage differs from the median coverage, indicating birth dose coverage is skewed and not normally distributed. The national Hepatitis B vaccine birth dose coverage overall was 72.4% in 2014 (National Immunization Survey). These data demonstrate clinically meaningful differences in performance affecting a large patient population (all newborns) for which continued improvement is warranted, and could help to justify not excluding parent refusals from the denominator, thereby making the measure more robust.

2b6. COMPARABILITY OF PERFORMANCE SCORES WHEN MORE THAN ONE SET OF SPECIFICATIONS
If only one set of specifications, this section can be skipped.

Note: *This item is directed to measures that are risk-adjusted (with or without SDS factors)* ***OR*** *to measures with more than one set of specifications/instructions (e.g., one set of specifications for how to identify and compute the measure from medical record abstraction and a different set of specifications for claims or eMeasures). It does not apply to measures that use more than one source of data in one set of specifications/instructions (e.g., claims data to identify the denominator and medical record abstraction for the numerator).* ***Comparability is not required when comparing performance scores with and without SDS factors in the risk adjustment model. However, if comparability is not demonstrated for measures with more than one set of specifications/instructions, the different specifications (e.g., for medical records vs. claims) should be submitted as separate measures.***

2b6.1. Describe the method of testing conducted to compare performance scores for the same entities across the different data sources/specifications (*describe the steps—do not just name a method; what statistical analysis was used*)

N/A

2b6.2. What were the statistical results from testing comparability of performance scores for the same entities when using different data sources/specifications? (*e.g., correlation, rank order*)

N/A

2b6.3. What is your interpretation of the results in terms of the differences in performance measure scores for the same entities across the different data sources/specifications? (i.*e., what do the results mean and what are the norms for the test conducted*)

N/A

2b7. MISSING DATA ANALYSIS AND MINIMIZING BIAS

2b7.1. Describe the method of testing conducted to identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias (*describe the steps—do not just name a method; what statistical analysis was used*)

By definition, data missing regarding administration of the Hepatitis B vaccine birth dose correspond to no birth dose administration, as 'negatives' (i.e., lack of vaccination) are not or may not be billed for, coded, or documented.

2b7.2. What is the overall frequency of missing data, the distribution of missing data across providers, and the results from testing related to missing data? (*e.g., results of sensitivity analysis of the effect of various rules for missing data/nonresponse; if no empirical sensitivity analysis, identify the approaches for handling missing data that were considered and pros and cons of each*)

N/A

2b7.3. What is your interpretation of the results in terms of demonstrating that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias**?** (i.*e., what do the results mean in terms of supporting the selected approach for missing data and what are the norms for the test conducted; if no empirical analysis, provide rationale for the selected approach for missing data*)

N/A

3. Feasibility

Extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

3a. Byproduct of Care Processes

For clinical measures, the required data elements are routinely generated and used during care delivery (e.g., blood pressure, lab test, diagnosis, medication order).

3a.1. **Data Elements Generated as Byproduct of Care Processes.**

Coded by someone other than person obtaining original information (e.g., DRG, ICD-9 codes on claims), Abstracted from a record by someone other than person obtaining original information (e.g., chart abstraction for quality measure or registry)

If other:

3b. Electronic Sources

The required data elements are available in electronic health records or other electronic sources. If the required data are not in electronic health records or existing electronic sources, a credible, near-term path to electronic collection is specified.

3b.1. **To what extent are the specified data elements available electronically in defined fields?** *(i.e., data elements that are needed to compute the performance measure score are in defined, computer-readable fields)*

Some data elements are in defined fields in electronic sources

3b.2. **If ALL the data elements needed to compute the performance measure score are not from electronic sources, specify a credible, near-term path to electronic capture, OR provide a rationale for using other than electronic sources.**

Not all hospitals/birthing facilities currently have the infrastructure for electronic sources, therefore paper sources of medical records, pharmacy, etc. must be used until such time as all hospital/birthing facilities have electronic source capabilities. However, adoption of electronic health records has increased; in 2014, 75.5% of acute care hospitals had adopted at least a basic electronic health record, representing an eight-fold increase from 2008.

3b.3. **If this is an eMeasure, provide a summary of the feasibility assessment in an attached file or make available at a measure-specific URL.**

Attachment **Attachment:** 1a.3..docx

3c. Data Collection Strategy

Demonstration that the data collection strategy (e.g., source, timing, frequency, sampling, patient confidentiality, costs associated with fees/licensing of proprietary measures) can be implemented (e.g., already in operational use, or testing demonstrates that it is ready to put into operational use). For eMeasures, a feasibility assessment addresses the data elements and measure logic and demonstrates the eMeasure can be implemented or feasibility concerns can be adequately addressed.

3c.1. **Describe what you have learned/modified as a result of testing and/or operational use of the measure regarding data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, time and cost of data collection, other feasibility/implementation issues.**

IF a PRO-PM, consider implications for both individuals providing PROM data (patients, service recipients, respondents) and those whose performance is being measured.

o The Feasibility Study was conducted in 50 hospitals using 2008 data that were representative of labor and delivery hospitals in Texas. The annual birth cohort size at study hospitals ranged from 219 to 6,530 and together represented more than 100,000 births. Hospitals were divided among urban and rural locations and included for-profit, not-for-profit, and public facilities. Thirty-six of 50 hospitals (72%), representing more than 62,000 births, calculated the measure. No issue of patient confidentiality was encountered. Chart reviews were done under the auspices of the public health authority of the Texas Department of State Health Services.

o Among 50 hospitals participating in the survey, overall 38 (76%) indicated they were able to provide data for the measure; 2 of these hospitals eventually did not provide the data. However, only 19 (38%) of the hospitals indicated they had access to data to calculate the number of parent/guardian vaccination refusals. The two most common reasons for not providing data were the time burden (71%) and information management (64%). With increasing uptake of electronic health records, these reasons may become less important over time.

o The cost of providing the measure was based on responses from hospitals participating in the Feasibility Study. None had previous experience providing the measure information, and thus, reflect a "start-up" cost.

o To determine the direct cost associated with determining the number of infants vaccinated with Hepatitis B vaccine prior to discharge, 6 hospitals provided information: mean $65, median $25, minimum $0, maximum $240 (2008 USD).

o To determine the indirect cost associated with determining the number of infants vaccinated with Hepatitis B vaccine prior to discharge, 11 hospitals provided information: mean $303, median $100, minimum $0, maximum $1650 (2008 USD).
o To determine the direct cost associated with determining the parent/guardian vaccination refusal rate (done before implementation of ICD-10 coding), 5 hospitals provided information: mean $594, median $10, minimum $0, maximum $2000 (2008 USD).
o To determine the indirect cost associated with determining the parent/guardian vaccination refusal rate, 6 hospitals provided information: mean $136, median $27, minimum $0, maximum $725 (2008 USD).
o Costs varied considerably by retrieval method:
  o The highest cost was associated with retrieving information from an electronic health record; 5 hospitals provided information: mean $970, median $1160, minimum $0, maximum $2000 (2008 USD). Presumably some or most of this cost entailed initial programming which might not be necessary in subsequent years.
  o The lowest cost was associated with retrieving information from an unknown source; 2 hospitals provided information.
 o The measure would be more robust if parent/guardian refusals were not excluded from the denominator.

3c.2. Describe any fees, licensing, or other requirements to use any aspect of the measure as specified *(e.g., value/code set, risk model, programming code, algorithm)*.
N/A

4. Usability and Use

Extent to which potential audiences (e.g., consumers, purchasers, providers, policy makers) are using or could use performance results for both accountability and performance improvement to achieve the goal of high-quality, efficient healthcare for individuals or populations.

4a. Accountability and Transparency

Performance results are used in at least one accountability application within three years after initial endorsement and are publicly reported within six years after initial endorsement (or the data on performance results are available). If not in use at the time of initial endorsement, then a credible plan for implementation within the specified timeframes is provided.

4.1. Current and Planned Use

NQF-endorsed measures are expected to be used in at least one accountability application within 3 years and publicly reported within 6 years of initial endorsement in addition to performance improvement.

Planned	Current Use (for current use provide URL)
Use Unknown	Public Reporting New York City Department of Health and Mental Hygiene http://www.nyc.gov/html/doh/html/hcp/cd-hepatitisb-pregnancy.shtml Public Health/Disease Surveillance National Immunization Survey http://www.cdc.gov/nchs/nis.htm Professional Certification or Recognition Program Immunization Action Coalition http://www.immunize.org/honor-roll/birthdose/ Quality Improvement with Benchmarking (external benchmarking to multiple organizations) Immunization Action Coalition http://www.immunize.org/honor-roll/birthdose/ Quality Improvement (Internal to the specific organization) 3 hospitals in New York City N/A

4a.1. For each CURRENT use, checked above, provide:

- Name of program and sponsor
- Purpose
- Geographic area and number and percentage of accountable entities and patients included

a. New York City Department of Health and Mental Hygiene; to offer healthcare professionals current information about prevention of perinatal transmission of Hepatitis B; New York City - 40 delivery facilities, 118,995 births
b. National Immunization Survey; to monitor childhood immunization coverage; 50 states, the District of Columbia, and some U.S. territories; 14,893 children (providing national estimate)
e. and f. Immunization Action Coalition Birth Dose Honor Roll; to recognize hospitals and birthing centers that have attained high coverage rates for administering Hepatitis B vaccine at birth; 220 hospitals in the United States
g. Hospitals in New York City and Michigan; to share results with nursery staff/obstetric providers/pediatric providers, used by quality assurance/risk management, reported to public health, publish in hospital newsletter/magazine/ website/promotional materials, used for research; 3 hospitals in New York City

Note that some uses do not exclude parent/guardian refusals from the denominator.

4a.2. If not currently publicly reported OR used in at least one other accountability application (e.g., payment program, certification, licensing) what are the reasons? (*e.g., Do policies or actions of the developer/steward or accountable entities restrict access to performance results or impede implementation?*)

N/A

4a.3. **If not currently publicly reported OR used in at least one other accountability application, provide a credible plan for implementation within the expected timeframes -- any accountability application within 3 years and publicly reported within 6 years of initial endorsement.** (*Credible plan includes the specific program, purpose, intended audience, and timeline for implementing the measure within the specified timeframes. A plan for accountability applications addresses mechanisms for data aggregation and reporting.*)
N/A

4b. Improvement

Progress toward achieving the goal of high-quality, efficient healthcare for individuals or populations is demonstrated. If not in use for performance improvement at the time of initial endorsement, then a credible rationale describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.

4b.1. **Progress on Improvement. (Not required for initial endorsement unless available.)**
Performance results on this measure (current and over time) should be provided in 1b.2 and 1b.4. Discuss:

- **Progress (trends in performance results, number and percentage of people receiving high-quality healthcare)**
- **Geographic area and number and percentage of accountable entities and patients included**

Data from the National Immunization Survey (providing national vaccine coverage estimates based upon 14,893 children with adequate data) demonstrate some improvement in Hepatitis B birth dose coverage, from 64.1% (+/-1.3) in 2010 to 72.4% (+/-1.5) in 2014. Similarly, the number of Immunization Action Coalition birth dose honorees (reflecting hospitals or birthing facilities which achieve a Hepatitis B vaccine birth dose coverage of 90% or greater) increased from 1 honoree in July 2013 to 210 honorees in September 2015.

4b.2. **If no improvement was demonstrated, what are the reasons? If not in use for performance improvement at the time of initial endorsement, provide a credible rationale that describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.**
N/A

4c. Unintended Consequences

The benefits of the performance measure in facilitating progress toward achieving high-quality, efficient healthcare for individuals or populations outweigh evidence of unintended negative consequences to individuals or populations (if such evidence exists).

4c.1. **Were any unintended negative consequences to individuals or populations identified during testing; OR has evidence of unintended negative consequences to individuals or populations been reported since implementation? If so, identify the negative unintended consequences and describe how benefits outweigh them or actions taken to mitigate them.**
Hospitals have not had a requirement to report data on Hepatitis B birth dose vaccination rates and/or parent/guardian refusals. As a consequence, the Feasibility Study demonstrated a wide variety of hospital capacity for providing the data for this measure. Some hospitals possessed the full capacity to produce the measure via easily accessible electronic or paper records. Others required laborious review of paper records. Other hospitals did not capture all the required information for the complete calculation in either a paper or electronic form, or kept some data electronically and some in paper records. Despite these challenges, most hospitals were able to provide the measure at a value within 10% of that determined by the sample of medical charts reviewed. A few hospitals provided a value that was considerably different. Inaccuracies in the calculations will be directly related to a given hospital´s information management practices for the data required. Although paper records will most likely require more time for review, this may not present an accuracy problem in hospitals with smaller delivery volumes that keep paper records and maintain them on site for easy access. Likewise, an electronic record management system will be accurate only so far as the data are required to be entered into the system, and the retrieval of the data is subsequently easy. ICD-9 and ICD-10 codes are available for both live births (V27.x) and hepatitis B vaccination (V05.3). Parent/guardian refusal of vaccination can be found in ICD-10 codes. The accuracy of the calculation excluding refusals will likely be adversely affected until hospitals develop effective information management of these data using these codes. An unintended consequence of reporting the measure after removing parent/guardian refusals will be the loss of information about hospitals with an unusually high refusal rates. Conversely, a hospital with high rates of vaccination might be targeted with anti-vaccine advocacy pressure.

5. Comparison to Related or Competing Measures

If a measure meets the above criteria and there are endorsed or new related measures (either the same measure focus or the same target population) or competing measures (both the same measure focus and the same target population), the measures are compared to address harmonization and/or selection of the best measure.

5. Relation to Other NQF-endorsed Measures
Are there related measures (conceptually, either same measure focus or target population) or competing measures (conceptually both the same measure focus and same target population)? If yes, list the NQF # and title of all related and/or competing measures.
No

5.1a. **List of related or competing measures (selected from NQF-endorsed measures)**

5.1b. **If related or competing measures are not NQF endorsed please indicate measure title and steward.**

5a. Harmonization

The measure specifications are harmonized with related measures;
OR
The differences in specifications are justified

5a.1. **If this measure conceptually addresses EITHER the same measure focus OR the same target population as NQF-endorsed measure(s):**
Are the measure specifications completely harmonized?

5a.2. **If the measure specifications are not completely harmonized, identify the differences, rationale, and impact on interpretability and data collection burden.**
N/A

5b. Competing Measures

The measure is superior to competing measures (e.g., is a more valid or efficient way to measure);
OR
Multiple measures are justified.

5b.1. **If this measure conceptually addresses both the same measure focus and the same target population as NQF-endorsed measure(s):**
Describe why this measure is superior to competing measures (e.g., a more valid or efficient way to measure quality); OR provide a rationale for the additive value of endorsing an additional measure. (Provide analyses when possible.)
N/A

Appendix

A.1 Supplemental materials may be provided in an appendix. All supplemental materials (such as data collection instrument or methodology reports) should be organized in one file with a table of contents or bookmarks. If material pertains to a specific submission form number, that should be indicated. Requested information should be provided in the submission form and required attachments. There is no guarantee that supplemental materials will be reviewed.
Attachment **Attachment:** FeasibilityReportTexas-635882735966472669.pdf

Contact Information

Co.1 **Measure Steward (Intellectual Property Owner):** Centers for Disease Control and Prevention
Co.2 **Point of Contact:** Sarah, Schillie, sschillie@cdc.gov, 404-718-8608-
Co.3 **Measure Developer if different from Measure Steward:** Centers for Disease Control and Prevention
Co.4 **Point of Contact:** Sarah, Schillie, sschillie@cdc.gov, 404-718-8608-

Additional Information

Ad.1 **Workgroup/Expert Panel involved in measure development**
Provide a list of sponsoring organizations and workgroup/panel members' names and organizations. Describe the members' role in measure development.
N/A

Measure Developer/Steward Updates and Ongoing Maintenance
Ad.2 **Year the measure was first released:** 2007
Ad.3 **Month and Year of most recent revision:** 10, 2011
Ad.4 **What is your frequency for review/update of this measure?**
Ad.5 **When is the next scheduled review/update for this measure?** 05, 2016

Ad.6 **Copyright statement:** I was an employee of the U.S. Federal Government when this work was conducted and prepared for publication; therefore, it is not protected by the Copyright Act, and copyright ownership cannot be transferred.
Ad.7 **Disclaimers:** The findings and conclusions in this report are those of the author(s) and do not necessarily represent the views of the Centers for Disease Control and Prevention.

Ad.8 **Additional Information/Comments:** Please note that the figure for 1.a.3 is also attached in a separate file (in case it is unreadable in the submission). Note that this separate file is attached at 3.b.3, as the Feasibility Report was attached for extra files (and hence 3.b.3 was the ´space´ that allowed for attachment of an additional file [Appendix]).

MEASURE WORKSHEET

This document summarizes the evaluation of the measure as it progresses through NQF's Consensus Development Process (CDP). The information submitted by measure developers/stewards is included after the Brief Measure Information, Preliminary Analysis, and Pre-meeting Public and Member Comments sections.

To navigate the links in the worksheet: Ctrl + click link to go to the link; ALT + LEFT ARROW to return

Brief Measure Information

NQF #: 0476
Measure Title: PC-03 Antenatal Steroids
Measure Steward: The Joint Commission
Brief Description of Measure: This measure assesses patients at risk of preterm delivery at >=24 and <34 weeks gestation receiving antenatal steroids prior to delivering preterm newborns. This measure is a part of a set of five nationally implemented measures that address perinatal care (PC-01: Elective Delivery, PC-02: Cesarean Birth, PC-04: Health Care-Associated Bloodstream Infections in Newborns, PC-05: Exclusive Breast Milk Feeding).
Developer Rationale: Antenatal corticosteroids are indicated for women at risk of premature delivery with few exceptions and will result in a substantial decrease in neonatal morbidity and mortality, as well as substantial savings in health care costs. The use of antenatal corticosteroids for fetal maturation is a rare example of a technology that yields substantial cost savings in addition to improving health. The Royal College of Obstetricians and Gynaecologists (RCOG) calculated that in increase in use from 15% to 60% in newborns with a birth weight less than 2000 gms born in the US would result in an annual savings of $157 million.

The measure will assist health care organizations (HCOs) to track evidence of an increase in the appropriate use of antenatal steroids prior to preterm deliveries.

Numerator Statement: Patients with antenatal steroids initiated prior to delivering preterm newborns (refer to Appendix C, Table 11.0, antenatal steroid medications available at: http://manual.jointcommission.org/releases/TJC2015B2/)
Denominator Statement: Patients delivering live preterm newborns with >=24 and <34 weeks gestation completed with ICD-10-PCS Principal or Other Procedure Codes for delivery as defined in Appendix A, Table 11.01.1 available at: http://manual.jointcommission.org/releases/TJC2015B2/
Denominator Exclusions: • Less than 8 years of age
• Greater than or equal to 65 years of age
• Length of Stay >120 days
• Enrolled in clinical trials
• Documented Reason for Not Initiating Antenatal Steroids
• ICD-10-CM Principal Diagnosis Code or ICD-10-CM Other Diagnosis Codes for fetal demise as defined in Appendix A, Table 11.09.1 available at: http://manual.jointcommission.org
• Gestational Age < 24 or >= 34 weeks or UTD

Measure Type: Process
Data Source: Paper Medical Records
Level of Analysis: Facility, Population : National

IF Endorsement Maintenance – Original Endorsement Date: Oct 24, 2008 **Most Recent Endorsement Date:** Apr 02, 2012

Maintenance of Endorsement -- Preliminary Analysis

To maintain NQF endorsement endorsed measures are evaluated periodically to ensure that the measures still meets the NQF endorsement criteria ("maintenance"). The emphasis for maintaining endorsement is focused on how effective the measure is for promoting improvements in quality. Endorsed measures should have some experience from the field to

inform the evaluation. The emphasis for maintaining endorsement is noted for each criterion.

Criteria 1: Importance to Measure and Report

1a. Evidence
Maintenance measures – less emphasis on evidence unless there is new information or change in evidence since the prior evaluation.

1a. Evidence. The evidence requirements for a *process or intermediate outcome* measure is that it is based on a systematic review (SR) and grading of the body of empirical evidence where the specific focus of the evidence matches what is being measured.

The developer provides the following evidence for this measure:

- **Systematic Review of the evidence specific to this measure?** ☒ **Yes** ☐ **No**
- **Quality, Quantity and Consistency of evidence provided?** ☒ **Yes** ☐ **No**
- **Evidence graded?** ☒ **Yes** ☐ **No**

Evidence Summary or **Summary of prior review in 2012:**

The evidence supporting this measure are 1) guidelines from the NIH Consensus Repot (1995/updated 2000) and 2) a 2012 Cochrane systematic review. The systematic reviews agreed and demonstrated a benefit for the use of antenatal steroids in reducing neonatal death, NICU admissions and multiple complications. ” The NIH consensus report included graded evidence for multiple endpoints relating to the benefits of antenatal corticosteroids. All evidence was Grade I with strength of recommendation for use for nearly all outcomes evaluated A or B. There were no outcomes where the strength of the recommendation was D or E.”

Changes to evidence from last review

☒ **The developer attests that there have been no changes in the evidence since the measure was last evaluated.**
☐ **The developer provided updated evidence for this measure:**

Updates: In the specifications section the developer reports that the measure has been changed to reflect the 2013 ACOG Practice Bulletin for Premature Rupture of Membranes that recommends antenatal steroids up to 34 weeks (change from 32 weeks).

Exception to evidence NA

Guidance from the Evidence Algorithm

Process measure (Box 1) → Systematic review (Box 3) → QQC present (Box 4) →SR concludes high quality evidence

Questions for the Committee:

- *The developer attests the underlying evidence for the measure has not changed since the last NQF endorsement review. Does the Committee agree the evidence basis for the measure has not changed and there is no need for repeat discussion and voting on Evidence?*

Preliminary rating for evidence: ☒ **Pass** ☐ **No Pass**

1b. Gap in Care/Opportunity for Improvement and 1b. Disparities
Maintenance measures – increased emphasis on gap and variation

1b. Performance Gap. The performance gap requirements include demonstrating quality problems and opportunity for improvement.

The developer provides initial data for 2Q 2010: 114 hospitals, 1225 patients; average measure rate – 63.3%

In January 2014 the measure became mandatory for all hospitals with >1100 births/year:

	2011	2012	2013	2014
# hospitals	115	122	146	1133
# patients	1351	1157	1436	13,343
National aggregate rate	73.6%	66.4%	88.5%	91.6%
10^{th}-25^{th} -90^{th} %tile	0 -0– 100%	0 – 40.0- 100%	45.3 – 58.6– 100%	37.8 -52.3– 100%
Mean hospital rate (SD)	0.54 (0.40)	0.65 (0.39)	0.73 (0.38)	0.82(0.31)

Disparities

The developer provides literature references rather than data from use of this measure:

- A 2011 report on births in California found that Hispanic mothers (25.6%), mothers younger than age 20 (27.6%), and those without prenatal care (52.2%) were less likely to receive antenatal steroids. Mothers giving birth vaginally (26.8%) and mothers with a diagnosis of fetal distress (26.5%) were also less likely to receive antenatal steroids.
- A 2015 a case control study at 24 university hospitals in the U.S. found that mothers of African American descent were less likely to have more than one prenatal visit and receive antenatal steroids prior to a premature delivery than White mothers.

Questions for the Committee:

- *Is there a gap in care that warrants a national performance measure?*
- *Can this measure be used to identify and track disparities of care?*

Preliminary rating for opportunity for improvement: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 1: Importance to Measure and Report (including 1a, 1b, 1c)

1a. Evidence to Support Measure Focus

Comments:

Process measure is directly related to the outcome by prevention.

1b. Performance Gap

Comments:

Yes, there is performance data. It shows gaps in care related to delivery mode, race, and number of prenatal care visits.

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability

2a1. Reliability Specifications

Maintenance measures – no change in emphasis – specifications should be evaluated the same as with new measures

2a1. Specifications requires the measure, as specified, to produce consistent (reliable) and credible (valid) results about the quality of care when implemented.

Data source(s):

- Paper medical records
- Vital Records reports, delivery logs and clinical information systems were added as acceptable data sources to help hospitals identify all cases

Specifications:

- The specifications have been converted from ICD-9 to ICD-10. The goal was to convert ICD-9 to ICD-10

equivalent codes, consistent with the clinical intent of the original measure specifications.

- The initial patient population is now identified with ICD-10-PCS-Principal or Other Procedure Codes for delivery instead of diagnosis codes for pregnancy.
- The single numerator data element "Antenatal Steroids Administered" was changed to "Antenatal Steroids Initiated" to capture initiation of antenatal steroids instead of a full course.
- The denominator statement was changed from patients delivering live preterm newborns with >=24 and <32 weeks gestation completed to patients delivering live preterm newborns with >=24 and <34 weeks gestation based on the 2013 ACOG Practice Bulletin on Premature Rupture of Membranes (PROM).
- There are eight data elements in the denominator: Admission Date; Birthdate; Clinical Trial; Discharge Date; Gestational Age; ICD-10-CM Other Diagnosis Codes; ICD-10-CM Principal Diagnosis Code; Reason for Not Initiating Antenatal Steroids
- Sampling is allowed.
- A calculation algorithm is presented.

Questions for the Committee:

- *Are the changes to the numerator and denominator appropriate?*
- *Are all the data elements clearly defined? Are all appropriate codes included?*
- *Is the logic or calculation algorithm clear?*
- *Is it likely this measure can be consistently implemented?*

2a2. Reliability Testing Testing attachment
Maintenance measures – less emphasis if no new testing data provided

2a2. Reliability testing demonstrates if the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise enough to distinguish differences in performance across providers.

For maintenance measures, summarize the reliability testing from the prior review:

Inter-rater reliability (IRR) was performed by ORXY vendor reabstaraction for 108 hospitals comprising 13,279 records. IRR is an appropriate method of assessing data element reliability for chart abstraction. The agreement rate for the data element "Antenatal steroids administered" was 99.16%.

Describe any updates to testing none

SUMMARY OF TESTING

Reliability testing level ☐ **Measure score** ☒ **Data element** ☐ **Both**

Reliability testing performed with the data source and level of analysis indicated for this measure ☒ **Yes** ☐ **No**

Method(s) of reliability testing see above

Results of reliability testing see above Only % agreement; no statistical test such as Kappa or ICC.

Questions for the Committee:

- *The developer does not provide any additional testing for reliability since the last NQF endorsement review. Does the Committee agree there is no need for repeat discussion and voting on Evidence?*

Guidance from the Reliability Algorithm

Precise specifications (Box 1) → Empirical reliability testing (Box 2) → Data element testing (Box 8) → appropriate method- IRR (Box 9) → high or moderate confidence of reliability of numerator data element → moderate (highest rating possible)

Preliminary rating for reliability: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

2b. Validity
Maintenance measures – less emphasis if no new testing data provided

2b1. Validity: Specifications

2b1. Validity Specifications. This section should determine if the measure specifications are consistent with the evidence.

In the specifications section the developer reports that the measure has been changed to reflect the 2013 ACOG Practice Bulletin for Premature Rupture of Membranes that recommends antenatal steroids up to 34 weeks (change from 32 weeks).

Specifications consistent with evidence in 1a. ☒ **Yes** ☐ **Somewhat** ☐ **No**
Specification not completely consistent with evidence

Question for the Committee:
- *Are the specifications consistent with the evidence?*

2b2. Validity testing

2b2. Validity Testing should demonstrate the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality.

For maintenance measures, summarize the validity testing from the prior review: face validity only

Describe any updates to validity testing: new empirical validity testing of the measure score is provided

SUMMARY OF TESTING
Validity testing level ☒ **Measure score** ☐ **Data element testing against a gold standard** ☐ **Both**

Method of validity testing of the measure score:
- ☐ **Face validity only**
- ☒ **Empirical validity testing of the measure score**

Validity testing method:
Using one year of data from 2014-2015 for 162 hospitals and 390,903 patient records, the Spearman rank-order correlation (a nonparametric measure of association based on the ranks of the data values by measure PC-03 and hospitals) was used to correlate the results from this measure with other measures in the Joint Commission's perinatal set. The developer hypothesizes that hospitals that perform well on this measure will perform well on all measures in the set.

Validity testing results: The developer provides scatter plots and a correlation table and summarized the findings: " The correlation of PC-03 with the other PC measures in the PC measure set indicates that the correlations with two other PC measures are moderate and statistically significant."

Questions for the Committee:
- *The correlations among the five measures are relatively weak. Would you expect all of the measures to be strongly correlated?*
- *Is the test sample adequate to generalize for widespread implementation?*
- *Do the results demonstrate sufficient validity so that conclusions about quality can be made?*
- *Do you agree that the score from this measure as specified is an indicator of quality?*

2b3-2b7. Threats to Validity

2b3. Exclusions:

Frequency of exclusions were presented:

Data element	Overall %	Median%	Min - Max
Clinical Trials	0.07	0	0 – 6.59%
Reason for not initiating antenatal steroid therapy	1.35	0.2	0 – 71.17% (90th %tile = 0.54%)
Codes for fetal demise	0.009	0	0 – 0.034%
Gestational age <24 and ≥ 34 weeks	96	96	85 - 98%

Questions for the Committee:

- *Are the exclusions consistent with the evidence?*
- *Are any patients or patient groups inappropriately excluded from the measure?*
- *Are the exclusions/exceptions of sufficient frequency and variation across providers to be needed (and outweigh the data collection burden)?*

2b4. Risk adjustment: **Risk-adjustment method** ☒ **None** ☐ **Statistical model** ☐ **Stratification**

2b5. Meaningful difference *(can statistically significant and clinically/practically meaningful differences in performance measure scores can be identified)*:

The developer reports the following descriptive analysis of the measure results:
N = 162 hospitals Number of records = 390,903
Mean = 98%
Min = 85%
Percentile 10%: 94%
Percentile 25%: 97%
Median: 99%
Percentile 75%: 100%
Percentile 90%: 100%
Max = 100%

The Joint Commission's Target Analysis uses two methods: Given the national average for a performance measure, a target range is constructed. Using generalized linear mixed models methodology (also known as hierarchical models), a predicted estimate of an HCO's performance, with a corresponding 95% confidence interval, is generated. This confidence interval is compared to the target range, to determine the HCOs' rating. The estimate of the organization's true performance is based on both the data from that organization and on data from the entire set of reporting organizations. The object of target analysis is to compare a health care organizations (HCO) data against a comparative norm for the purpose of evaluating performance improvement opportunities. When an organization's performance level is statistically significantly different from a comparative norm, it is considered a statistical deviation. A statistical deviation may be desirable or undesirable depending on the "direction of improvement" of the measure.

PC-03 Distribution of Outliers 2011 1st Quarter Data:
Scores on this measure: N=79, Mean 64.6%, SD 0.40336
10th Percentile= 0%
25th Percentile= 25%
50th Percentile= 80%
75th Percentile= 100%
90th Percentile= 100%

79 (100%) Neutral – results not significantly different from target range

Question for the Committee:

- *If the performance for all hospitals is not significantly different from the target range, what does this measure tell us*

about meaningful differences about quality?

2b6. Comparability of data sources/methods: NA

2b7. Missing Data
The developer reports that "Any file with missing data will result in a measure category assignment of X and rejection of the file from the warehouse, unless the data are not used to process the measure. If the data are used to process the measure and have been reported by the abstractor as No/UTD, the case will result in a measure category assignment of D (i.e., failed measure, not rejected)."

Guidance from algorithm: Specifications consistent with evidence (Box 1) → Assessed all potential threats to validity (Box 2) **Meaningful differences may be an issue here** → empirical testing (Box 3) →testing of measure score (Box 6) →appropriate testing method (Box 7) → testing results moderate (or possible low) → moderate

Preliminary rating for reliability: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**
If lack of meaningful differences and results of empirical validity testing are a concern, then rating may be more appropriately LOW.

Committee pre-evaluation comments

Criteria 2: Scientific Acceptability of Measure Properties (including all 2a, 2b, and 2d)

2a1. & 2b1. Specifications
Comments:
The specifications are consistent with the evidence

2a2. Reliability Testing
Comments:
Adequate, no need for further testing.

2b2. Validity Testing
Comments:
Adequate scope. Empirical validity testing did occur. Sufficient sample size with improved outcomes with steroid administration.
2b3. Exclusions Analysis
2b4. Risk Adjustment/Stratification for Outcome or Resource Use Measures
2b5. Identification of Statistically Significant & Meaningful Differences In Performance
2b6. Comparability of Performance Scores When More Than One Set of Specifications
2b7. Missing Data Analysis and Minimizing Bias
Comments:
no, sufficient sample size.

Criterion 3. Feasibility
Maintenance measures – no change in emphasis – implementation issues may be more prominent

3. Feasibility is the extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

- According to the developer "hospitals using this performance measure generally collect measure data via manual review of the paper medical record, the EMR or a combination of both."

Questions for the Committee:

- *What is the burden for hospitals in collecting data for this measure?*
- *Are the required data elements routinely generated and used during care delivery?*

○ *Are the required data elements available in electronic form, e.g., EHR or other electronic sources?*
○ *Is the data collection strategy ready to be put into operational use?*

Preliminary rating for feasibility: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 3: Feasibility

3a. Byproduct of Care Processes
3b. Electronic Sources
3c. Data Collection Strategy
Comments:
None. Should be available in electronic format. No concerns about operationalization.

Criterion 4: Usability and Use
Maintenance measures – increased emphasis – much greater focus on measure use and usefulness, including both impact /improvement and unintended consequences

4. Usability and Use evaluate the extent to which audiences (e.g., consumers, purchasers, providers, policymakers) use or could use performance results for both accountability and performance improvement activities.

Current uses of the measure
Publicly reported? ☒ **Yes** ☐ **No**

Current use in an accountability program? ☒ **Yes** ☐ **No**
OR
Planned use in an accountability program? ☐ **Yes** ☐ **No**

Accountability program details
- Public Reporting --Quality Check® http://www.qualitycheck.org/consumer/searchQCR.aspx
- Regulatory and Accreditation Programs Hospital Accreditation Program http://jointcommission.org

Improvement results
The developer summarizes "The rate of patients receiving antenatal steroids prior to premature deliveries has increased dramatically from 63.3% in 2010 with 114 hospitals reporting to 91.6% in 2014 with 1133 hospitals reporting based on Joint Commission OPRYX performance measurement data from across the country."

Unexpected findings (positive or negative) during implementation
The developers report on three unexpected findings during implementation of this measure:
- Cases were inappropriately failing when the repeat dose of antenatal steroids was not given due to the delivery occurring prior to the routinely scheduled repeat dose being ordered at the time the first dose was given.
  - Mitigating Action: The data element Antenatal Steroids Administered was changed to Antenatal Steroids Initiated to capture initiation of antenatal steroids instead of a full course.
- Patients who did not receive prenatal care were inappropriately included in the measure denominator, as the gestational age data element was abstracted as unable to be determined (UTD).
  - Mitigating Action: In order to avoid penalizing hospitals, cases with UTD were removed from the measure population.
- Some hospitals have reported lower rates due to small denominator populations as a result of sampling.

- Mitigating Action: Vital Records reports, delivery logs and clinical information systems were added as acceptable data sources to help hospitals identify all cases with 24 and less than 34 weeks gestation, so that 100% of these cases could be reviewed to increase the denominator population size.

Potential harms none reported

Feedback: no specific feedback on this measure

***Questions for the Committee**:*

- *Committee members who are using /have used this measure should share their experiences. Are you aware of any unintended consequences of this measure?*
- *How can the performance results be used to further the goal of high-quality, efficient healthcare?*
- *Do the benefits of the measure outweigh the burden of data collection or any potential unintended consequences?*

Preliminary rating for usability and use: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 4: Usability and Use

4a. Accountability and Transparency
4b. Improvement
4c. Unintended Consequences
Comments:
**TJC measure. Public Reporting-Quality Check. Unaware of any unintended consequences. **

Criterion 5: Related and Competing Measures

Related or competing measures
None

Harmonization NA

Pre-meeting public and member comments

•

1. IMPACT, OPPORTUITY, EVIDENCE - IMPORTANCE TO MEASURE AND REPORT

Importance to Measure and Report is a threshold criterion that must be met in order to recommend a measure for endorsement. All three subcriteria must be met to pass this criterion. See guidance on evidence.

Measures must be judged to be important to measure and report in order to be evaluated against the remaining criteria. (evaluation criteria)

1c.1 Structure-Process-Outcome Relationship *(Briefly state the measure focus, e.g., health outcome, intermediate clinical outcome, process, structure; then identify the appropriate links, e.g., structure-process-health outcome; process- health outcome; intermediate clinical outcome-health outcome)*:
The focus of the measure is to increase the appropriate use of antenatal steroids in preterm deliveries >> population determined >> population assessed >> antenatal steroids administered >> improved fetal lung maturity >> decreased fetal morbidity and mortality.

1c.2-3 Type of Evidence *(Check all that apply)*:
Clinical Practice Guideline
Systematic review of body of evidence (other than within guideline development)

1c.4 Directness of Evidence to the Specified Measure *(State the central topic, population, and outcomes addressed in the body of evidence and identify any differences from the measure focus and measure target population)*:
The central topic for the measure is the appropriate administration of a full course of antenatal steroids prior to a live preterm delivery at 24 0/7 to 32 0/7 weeks gestation. The target population for the performance measure is consistent with the body of evidence supporting the use of antenatal steroids for preterm deliveries.

1c.5 Quantity of Studies in the Body of Evidence *(Total number of studies, not articles)*: The Cochrane review updated and published in 2010 included 21 randomized control trials with nearly 3,900 women and more than 4,200 infants.

1c.6 Quality of Body of Evidence (*Summarize the certainty or confidence in the estimates of benefits and harms to patients across studies in the body of evidence resulting from study factors. Please address: a) study design/flaws; b) directness/indirectness of the evidence to this measure (e.g., interventions, comparisons, outcomes assessed, population included in the evidence); and c) imprecision/wide confidence intervals due to few patients or events):* The quality of evidence supporting the appropriate use of antenatal steroids is high as noted in the recent Cochrane review published in 2010. All evidence was Grade I with strength of recommendation for use for nearly all outcomes evaluated A or B, i.e., significant reductions in neonatal death, (RR 0.69), respiratory distress syndrome (RDS) (RR 0.66), intraventricular hemorrhage (IVH) (RR 0.54), necrotizing enterocolitis (NEC) (RR 0.46), early onset infection (RR 0.56), and NICU admission (RR 0.80). There were no outcomes where the strength of the recommendation was D or E. No study design flaws were noted. In all studies, a complete course of antenatal steroids compared to a placebo demonstrated accelerated fetal lung maturation supporting the continued use of antenatal steroids for preterm deliveries.

1c.7 Consistency of Results across Studies *(Summarize the consistency of the magnitude and direction of the effect):* The Cochrane review (2010) looks at multiple outcomes many of which are highlighted in 1.a.3. Although the results consistently showed benefit across studies, the confidence intervals are often wide for individual studies. Benefits include the significant reduction in neonatal death, (RR 0.69), respiratory distress syndrome (RDS) (RR 0.66), intraventricular hemorrhage (IVH) (RR 0.54), necrotizing enterocolitis (NEC) (RR 0.46), early onset infection (RR 0.56), and NICU admission (RR 0.80) were noted. There was no increase in maternal chorioamnionitis with a narrow confidence interval. There was also no increase in maternal mortality; however, the confidence interval is very wide due to the rarity of the event.

1c.8 Net Benefit *(Provide estimates of effect for benefit/outcome; identify harms addressed and estimates of effect; and net benefit - benefit over harms)*:
As described before, appropriate use of antenatal steroids consistently resulted in improved outcomes and a substantial decrease in neonatal morbidity and mortality, as well as substantial savings in health care costs The Royal College of Obstetricians and Gynaecologists (RCOG) calculated that in increase in use from 15% to 60% in babies of less than 2000 GM born in the US would result in an annual savings of $157 million.

1c.9 Grading of Strength/Quality of the Body of Evidence. Has the body of evidence been graded? Yes

1c.10 If body of evidence graded, identify the entity that graded the evidence including balance of representation and any disclosures regarding bias: US Preventive Services Task Force

1c.11 System Used for Grading the Body of Evidence: GRADE

1c.12 If other, identify and describe the grading scale with definitions:

1c.13 Grade Assigned to the Body of Evidence: The NIH consensus report included graded evidence for multiple endpoints relating to the benefits of antenatal corticosteroids. All evidence was Grade I with strength of recommendation for use for nearly all outcomes evaluated A or B. There were no outcomes where the strength of the recommendation was D or E.

1c.14 Summary of Controversy/Contradictory Evidence: There is no documented evidence regarding controversy related to appropriate use of antenatal steroids. Research reviewed many studies supporting the administration of a full course of antenatal steroids. The 1994 NIH consensus statement has remained the standard of care which was reaffirmed by NIH in 2000 and again in the recent Cochrane review published in 2010. There is some older literature suggesting antenatal steroids may not be effective in hypertensive patients, but this has not been borne out. There is basic science literature suggesting that hyperglycemia may inhibit surfactant action. Since steroids raise glucose levels this is a theoretic concern in diabetics, but there is no clinical evidence to alter practice. There are animal data and limited retrospective human data that higher doses have adverse effects on growth and development, but not at the doses currently used. There is long term follow-up as long as 30 years showing no adverse effects in humans given this regime.

1c.15 Citations for Evidence other than Guidelines*(Guidelines addressed below)*:

- National Institutes of Health (NIH) Consensus Conference. (1995). Effect of corticosteroids for fetal maturation on perinatal outcomes. JAMA. 273(5), 413-8.
- National Institutes of Health (HIH) Consensus Statement. (2000). Antenatal corticosteroids revisited: repeat courses. 17(2)1-18.
- Roberts, D. & Dalziel, S.R. (2010) Antenatal corticosteroids for accelerating fetal lung maturation for women at risk of preterm birth (Review). The Cochrane Collaboration. Issue 9.

1c.16 Quote verbatim, <u>the specific guideline recommendation</u> *(Including guideline # and/or page #)*:

Summary of Recommendations on Page 417: The benefits of antenatal administration of corticosteroids to fetuses at risk of preterm delivery vastly outweigh the potential risks. These benefits include not only a reduction in the risk of RDS but also a substantial reduction in mortality and IVH.

- All fetuses between 24 and 34 weeks´ gestation at risk of preterm delivery should be considered candidates for antenatal treatment with corticosteroids.
- The decision to use antenatal corticosteroids should not be altered by fetal race or gender or by the availability of surfactant replacement therapy.
- Patients eligible for therapy with tocolytics should also be eligible for treatment with antenatal corticosteroids.
- Treatment consists of two doses of 12 mg of betamethasone given intramuscularly 24 hours apart or four doses of 6 mg of dexamethasone given intramuscularly 12 hours apart. Optimal benefit begins 24 hours after initiation of therapy and lasts 7 days.
- Because treatment with corticosteroids for less than 24 hours is still associated with significant reductions in neonatal mortality, RDS, and IVH, antenatal corticosteroids should be given
unless immediate delivery is anticipated.
- In PPROM at less than 30 to 32 weeks´ gestation in the absence of clinical chorioamnionitis, antenatal corticosteroid use is recommended because of the high risk of IVH at these early gestational ages.
- In complicated pregnancies where delivery prior to 34 weeks´ gestation is likely, antenatal corticosteroid use is recommended unless there is evidence that corticosteroids will have an adverse effect on the mother or delivery is imminent.

1c.17 Clinical Practice Guideline Citation: • National Institutes of Health (NIH) Consensus Conference. (1995). Effect of corticosteroids for fetal maturation on perinatal outcomes. JAMA. 273(5), 413-8.

1c.18 National Guideline Clearinghouse or other URL:

1c.19 Grading of Strength of Guideline Recommendation. Has the recommendation been graded? Yes

1c.20 If guideline recommendation graded, identify the entity that graded the evidence including balance of representation

and any disclosures regarding bias: National Institutes of Health

1c.21 System Used for Grading the Strength of Guideline Recommendation: USPSTF

1c.22 If other, identify and describe the grading scale with definitions:

1c.23 Grade Assigned to the Recommendation: Level I

1c.24 Rationale for Using this Guideline Over Others: These guidelines have been developed for healthcare personnel by a nonfederal, nonadvocate, 16-member consensus panel including representatives from neonatology, obstetrics, family medicine, behavioral medicine, psychology, biostatistics, and the public; 19 experts in neonatology, obstetrics, and pharmacology presented data to the consensus panel and a conference audience of approximately 500. An extensive bibliography of references was produced for the consensus panel and the conference audience using a variety of on-line databases including MEDLINE. The consensus panel met several times prior to the conference to review the literature. It also commissioned an updated meta-analysis, a neonatal registry review, and an economic analysis that were presented at the conference. The experts prepared abstracts for distribution at the conference, presented data, and answered questions from the panel and audience. The panel evaluated the strength of the scientific evidence using the grading system developed by the Canadian Task Force on the Periodic Health Examination and adapted by the US Preventive Services Task Force. The guidelines continue to be affirmed by the American College of Obstetrics and Gynecology (ACOG) and the American Academy of Pediatrics (AAP) in their guidelines for perinatal care.

Based on the NQF descriptions for rating the evidence, what was the <u>developer's assessment</u> of the quantity, quality, and consistency of the body of evidence?

1c.25 Quantity: High **1c.26** Quality: High**1c.27** Consistency**:** High

1. Evidence, Performance Gap, Priority – Importance to Measure and Report

Extent to which the specific measure focus is evidence-based, important to making significant gains in healthcare quality, and improving health outcomes for a specific high-priority (high-impact) aspect of healthcare where there is variation in or overall less-than-optimal performance. ***Measures must be judged to meet all subcriteria to pass this criterion and be evaluated against the remaining criteria.***

1a. Evidence to Support the Measure Focus – See attached Evidence Submission Form
0476_Evidence_MSF5.0_Data-635787040704713660.doc

1b. Performance Gap

Demonstration of quality problems and opportunity for improvement, i.e., data demonstrating:

- considerable variation, or overall less-than-optimal performance, in the quality of care across providers; and/or
- disparities in care across population groups.

1b.1. **Briefly explain the rationale for this measure** *(e.g., the benefits or improvements in quality envisioned by use of this measure)*
Antenatal corticosteroids are indicated for women at risk of premature delivery with few exceptions and will result in a substantial decrease in neonatal morbidity and mortality, as well as substantial savings in health care costs. The use of antenatal corticosteroids for fetal maturation is a rare example of a technology that yields substantial cost savings in addition to improving health. The Royal College of Obstetricians and Gynaecologists (RCOG) calculated that in increase in use from 15% to 60% in newborns with a birth weight less than 2000 gms born in the US would result in an annual savings of $157 million.

The measure will assist health care organizations (HCOs) to track evidence of an increase in the appropriate use of antenatal steroids prior to preterm deliveries.

1b.2. **Provide performance scores on the measure as specified (current and over time) at the specified level of analysis.** *(This is required for endorsement maintenance. Include mean, std dev, min, max, interquartile range, scores by decile. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included). This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*
Administration of antenatal steroids prior to premature deliveries continues to improve; however, a performance gap still exists. Based on recommendations from the 1994 National Institutes of Health (NIH) consensus panel, a goal of 100% should be achievable. The Perinatal Care (PC) core measures were added as a new core measure set in 2010 for hospitals to select in order to meet their ORYX performance measurement requirement for Joint Commission accreditation purposes. At that time, approximately 114 hospitals reported the data with an average measure rate of 63.3% (n=1225 patients). In January 2014, The Joint Commission required mandatory reporting of the PC measure set for all accredited hospitals with 1100 births or more annually. 1133 hospitals reported the data with an average rate of 91.6% (n=13,343 patients). It is important to note that a performance gap of 62.6% exists for the 10th percentile of hospitals performing at 37.8% (if 100% is considered goal performance). Below is the specified level of analysis for PC-01 beginning with discharges April 1, 2010 through December 31, 2014.
2Q 2010: 1225 denominator cases; 776 numerator cases; 114 hospitals; 63.3% national aggregate rate; 0.53806 mean of hospital rates; 0.36806 standard deviation; 100% 90th percentile rate; 86.3% 75th percentile rate/upper quartile; 62.2% 50th percentile rate/median rate; 14.2% 25th percentile rate/lower quartile; and 0% 10th percentile rate.
CY 2011: 1351 denominator cases; 995 numerator cases; 115 hospitals; 73.6% national aggregate rate; 0.56437 mean of hospital rates; 0.39983 standard deviation; 100% 90th percentile rate; 100% 75th percentile rate/upper quartile; 68.7% 50th percentile rate/median rate; 0% 25th percentile rate/lower quartile; and 0% 10th percentile rate.
CY 2012: 1157 denominator cases; 947 numerator cases; 122 hospitals; 66.4% national aggregate rate; 0.66441 mean of hospital rates; 0.39206 standard deviation; 100% 90th percentile rate; 100% 75th percentile rate/upper quartile; 85.1% 50th percentile rate/median rate; 40.0% 25th percentile rate/lower quartile; and 0% 10th percentile rate.
CY 2013: 1436 denominator cases; 1271 numerator cases; 146 hospitals; 88.5% national aggregate rate; 0.73069 mean of hospital rates; 0.37887 standard deviation; 100% 90th percentile rate; 100%% 75th percentile rate/upper quartile; 99.0% 50th percentile rate/median rate; 58.6% 25th percentile rate/lower quartile; and 45.3% 10th percentile rate.
CY 2014: 13,343 denominator cases; 12,235 numerator cases; 1133 hospitals; 91.6% national aggregate rate; 0.82072 mean of hospital rates; 0.30696 standard deviation; 100% 90th percentile rate; 100% 75th percentile rate/upper quartile; 100% 50th percentile rate/median rate; 52.3% 25th percentile rate/lower quartile; and 37.8% 10th percentile rate.

1b.3. **If no or limited performance data on the measure as specified is reported in 1b2, then provide a summary of data from the literature that indicates opportunity for improvement or overall less than optimal performance on the specific focus of measurement.**

Not Applicable

1b.4. **Provide disparities data from the measure as specified (current and over time) by population group, e.g., by race/ethnicity, gender, age, insurance status, socioeconomic status, and/or disability.** *(This is required for endorsement maintenance. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities include.) This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*
No national data were available to report disparities; however, clinical data for premature newborns born in 2005-2007 in California for antenatal steroid administration for newborns with a birth weight of less than 1500 gms or gestational age less than 34 weeks was reviewed by the California Perinatal Quality Care Collaborative. They collect data on more than 90% of newborn admissions in California. Hispanic mothers (25.6%), mothers younger than age 20 (27.6%), and those without prenatal care (52.2%) were less likely to receive antenatal steroids. Mothers giving birth vaginally (26.8%) and mothers with a diagnosis of fetal distress (26.5%) were also less likely to receive antenatal steroids. After risk adjustment, the most prominent factors noted were neonatal level of care and lack of prenatal care (Lee et. al., 2011).
Shankaran, et al. (2015) conducted a case control study at 24 university hospitals in the U.S. and found that mothers of African American descent were less to have more than one prenatal visit and receive antenatal steroids prior to a premature delivery than White mothers. As a result, the risk of developing intraventricular hemorrhage (IVH) was higher among African American newborns. The authors concluded the risk for IVH differed between African American newborns and White newborns, possibly attributable to both race and health care disparities.

1b.5. **If no or limited data on disparities from the measure as specified is reported in 1b4, then provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement. Include citations.**
• Lee, H.C., Lyndon, A., Blumenfeld, Y.J., Dudley, R.A. & Gould, J.B. (2011). Antenatal steroid administration for premature neonates in California. Obstetrics & Gynecology, 117(3), 603-9.
• Shankaran, S., Lin, A., Maller-Kesselman, J., Zhang, H., O´Shea, T., et al. (2015). Maternal race, demography and health care disparities impact risk for IVH in preterm neonates. J Pediatr. 2014 May;164(5):1005-1011.e3. doi: 10.1016

1c. High Priority (previously referred to as High Impact)
The measure addresses:

- a specific national health goal/priority identified by DHHS or the National Priorities Partnership convened by NQF; OR
- a demonstrated high-priority (high-impact) aspect of healthcare (e.g., affects large numbers of patients and/or has a substantial impact for a smaller population; leading cause of morbidity/mortality; high resource use (current and/or future); severity of illness; and severity of patient/societal consequences of poor quality).

1c.1. **Demonstrated high priority aspect of healthcare**
A leading cause of morbidity/mortality, Patient/societal consequences of poor quality
1c.2. **If Other:**

1c.3. **Provide epidemiologic or resource use data that demonstrates the measure addresses a high priority aspect of healthcare. List citations in 1c.4.**
An updated systematic review by the Cochrane Collaboration analyzed 21 studies including 3885 women and 4269 infants. This review concluded that the reported use of antenatal corticosteroids resulted in significant reductions in neonatal death, (RR 0.69), respiratory distress syndrome (RDS) (RR 0.66), intraventricular hemorrhage (IVH) (RR 0.54), necrotizing enterocolitis (NEC) (RR 0.46), early onset infection (RR 0.56), and NICU admission (RR 0.80). There were no adverse maternal effects.

RDS affects up to one fifth of low birthweight babies defined as less than 2500 grams, and extremely low birthweight babies defined as less than 1500 grams. This deadly illness that preterm babies commonly suffer from is the primary cause of early neonatal death and disability (Roberts & Dalziel, 2010).

The evidence from this new systematic review supports the continued use of a single course of antenatal corticosteroids to accelerate fetal lung maturation in women at risk of preterm birth. This evidence greatly supports the use of a single course of antenatal corticosteroids to be considered routine for preterm delivery with few exceptions (Roberts & Dalziel, 2010).

1c.4. **Citations for data demonstrating high priority provided in 1a.3**
• American College of Obstetricians and Gynecologists (ACOG) Practice Bulletin No 139. (2013). Clinical Management Guidelines for Obstetrician-Gynecologists for Premature Rupture of Membranes.
• Falah, N. & Haas, D. (2014). Antenatal corticosteroid therapy: current strategies and identifying mediators and markers for

response. Semin Perinatol. 38(8):528-33. doi: 10.1053.
• Roberts, D. & Dalziel, S.R. (2010) Antenatal corticosteroids for accelerating fetal lung maturation for women at risk of preterm birth (Review). The Cochrane Collaboration. Issue 9.
• Sotiriadis, A., Tsiami, A., Papatheodorou, S., Baschatm A., Sarafidis, K. & Makrydimas, G. (2015). Neurodevelopmental outcome after a single course of antenatal steroids in children born preterm: a systematic review and meta-analysis. Obstet Gynecol. 125(6):1385-96. doi: 10.1097.

1c.5. **If a PRO-PM (e.g. HRQoL/functional status, symptom/burden, experience with care, health-related behaviors), provide evidence that the target population values the measured PRO and finds it meaningful.** *(Describe how and from whom their input was obtained.)*
Not Applicable

2. Reliability and Validity—Scientific Acceptability of Measure Properties

Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. ***Measures must be judged to meet the subcriteria for both reliability and validity to pass this criterion and be evaluated against the remaining criteria.***

2a.1. Specifications The measure is well defined and precisely specified so it can be implemented consistently within and across organizations and allows for comparability. eMeasures should be specified in the Health Quality Measures Format (HQMF) and the Quality Data Model (QDM).

De.5. **Subject/Topic Area** *(check all the areas that apply):*
Perinatal and Reproductive Health, Perinatal and Reproductive Health : Perinatal

De.6. **Cross Cutting Areas** *(check all the areas that apply):*
Disparities, Health and Functional Status, Health and Functional Status : Functional Status, Safety, Safety : Complications

S.1. **Measure-specific Web Page** *(Provide a URL link to a web page specific for this measure that contains current detailed specifications including code lists, risk model details, and supplemental materials. Do not enter a URL linking to a home page or to general information.)*
http://manual.jointcommission.org/releases/TJC2015B2/

S.2a. **If this is an eMeasure**, HQMF specifications must be attached. Attach the zipped output from the eMeasure authoring tool (MAT) - if the MAT was not used, contact staff. (Use the specification fields in this online form for the plain-language description of the specifications)
This is not an eMeasure **Attachment:**

S.2b. **Data Dictionary, Code Table, or Value Sets** *(and risk model codes and coefficients when applicable) must be attached. (Excel or csv file in the suggested format preferred - if not, contact staff)*
Attachment **Attachment:** PC03_ICD_Code_Tables.xlsx

S.3. **For endorsement maintenance**, please briefly describe any changes to the measure specifications since last endorsement date and explain the reasons.

- All ICD-9-CM diagnosis codes and ICD-9-CM procedure codes were converted to ICD-10-CM diagnosis and ICD-10-PCS procedure codes throughout the measure specifications. The Department of Health and Human Services (HHS) mandated that all entities covered by the Health Insurance Portability and Accountability Act (HIPAA) must transition to a new set of codes for electronic health care transactions on October 1, 2015
- The initial patient population is now identified with ICD-10-PCS-Principal or Other Procedure Codes for delivery instead of diagnosis codes for pregnancy, since the ICD-10-CM Principal or Other Diagnosis Codes do not indicate whether the delivery took place during the hospitalization.
- The numerator data element Antenatal Steroids Administered was changed to Antenatal Steroids Initiated to capture initiation of antenatal steroids instead of a full course. The change was made to reduce the burden of data abstraction, since a repeat dose was almost always not given due to the delivery occurring prior to the scheduled repeat dose.
- Cases with a gestational age of UTD were added to the denominator excluded populations, since UTD is highly correlated with no prenatal care.
- The denominator statement was changed from patients delivering live preterm newborns with >=24 and <32 weeks gestation completed to patients delivering live preterm newborns with >=24 and <34 weeks gestation based on the 2013 American College of Obstetricians & Gynecologists Practice Bulletin on Premature Rupture of Membranes (PROM) which recommends antenatal steroids for patients with PROM up to 34 weeks gestation instead of 32 weeks gestation.
- Vital Records reports, delivery logs and clinical information systems were added as acceptable data sources to help hospitals identify all cases with 24 and less than 34 weeks gestation, so that 100% of these cases could be reviewed to increase the denominator population size.

S.4. **Numerator Statement** *(Brief, narrative description of the measure focus or what is being measured about the target population, i.e., cases from the target population with the target process, condition, event, or outcome)*
IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
Patients with antenatal steroids initiated prior to delivering preterm newborns (refer to Appendix C, Table 11.0, antenatal steroid medications available at: http://manual.jointcommission.org/releases/TJC2015B2/)

S.5. **Time Period for Data** *(What is the time period in which data will be aggregated for the measure, e.g., 12 mo, 3 years, look back to August for flu vaccination? Note if there are different time periods for the numerator and denominator.)*
Episode of care

S.6. **Numerator Details** *(All information required to identify and calculate the cases from the target population with the target process, condition, event, or outcome such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
IF an OUTCOME MEASURE, describe how the observed outcome is identified/counted. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
One data element is used to calculate the numerator:

1. Antenatal Steroids Initiated- Documentation that antenatal steroids were initiated before delivery. Initial antenatal steroid therapy is 12mg betamethasone IM or 6mg dexamethasone IM. Allowable values: Yes or No/UTD. Cases are eligible for the numerator population when allowable value = Yes is selected.
Updates available at: http://manual.jointcommission.org/releases/TJC2015B2/

S.7. **Denominator Statement** *(Brief, narrative description of the target population being measured)*
Patients delivering live preterm newborns with >=24 and <34 weeks gestation completed with ICD-10-PCS Principal or Other Procedure Codes for delivery as defined in Appendix A, Table 11.01.1 available at: http://manual.jointcommission.org/releases/TJC2015B2/

S.8. **Target Population Category** *(Check all the populations for which the measure is specified and tested if any):*
Maternal Health

S.9. **Denominator Details** *(All information required to identify and calculate the target population/denominator such as definitions, specific data collection items/responses , code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
Eight data elements are used to calculate the denominator:

1. Admission Date – The month, day and year of admission to acute inpatient care.
2. Birthdate - The month, day and year the patient was born.
3. Clinical Trial - Documentation that during this hospital stay the patient was enrolled in a clinical trial in which patients with pregnancy were being studied. Allowable values: Yes or No/UTD
4. Discharge Date – The month day and year the patient was discharged from acute care, left against medical advice or expired during the stay.
5. Gestational Age – Documentation of the weeks of gestation completed at the time of delivery. Allowable Values: 1-50 or UTD.
6. ICD-10-CM Other Diagnosis Codes - The International Classification of Diseases, Tenth Revision, Clinical Modification codes associated with the secondary diagnoses for this hospitalization.
7. ICD-10-CM Principal Diagnosis Code - The International Classification of Diseases, Tenth Revision, Clinical Modification code associated with the diagnosis established after study to be chiefly responsible for occasioning the admission of the patient for this hospitalization.
8. Reason for Not Initiating Antenatal Steroids - Reasons for not initiating antenatal steroids before delivery are clearly documented in the medical record. Reasons for not initiating antenatal steroids may include fetal distress, imminent delivery or other reasons documented by physician/APN/PA/CNM. Allowable Values: Yes or No/UTD
Updates available at: http://manual.jointcommission.org/releases/TJC2015B2/

S.10. **Denominator Exclusions** *(Brief narrative description of exclusions from the target population)*
- Less than 8 years of age
- Greater than or equal to 65 years of age
- Length of Stay >120 days
- Enrolled in clinical trials
- Documented Reason for Not Initiating Antenatal Steroids
- ICD-10-CM Principal Diagnosis Code or ICD-10-CM Other Diagnosis Codes for fetal demise as defined in Appendix A, Table 11.09.1 available at: http://manual.jointcommission.org
- Gestational Age < 24 or >= 34 weeks or UTD

S.11. **Denominator Exclusion Details** *(All information required to identify and calculate exclusions from the denominator such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*

- The patient age in years is equal to the Admission Date minus the Birthdate. Patients less than 8 years of age or greater or equal to 65 years of age are excluded.
- Length of stay (LOS) in days is equal to the Discharge Date minus the Admission Date. If the LOS is greater than 120 days, the patient is excluded.
- Patients are excluded if "Yes" is selected for Clinical Trial.
- The data element Reason for Not Initiating Antenatal Steroids is used to determine if the patient had a documented reason for not receiving antenatal steroids.
- Patients with ICD-10-CM Principal Diagnosis Code or ICD-10-CM Other Diagnosis Codes for fetal demise are excluded.
- Patients with a Gestational Age less than 24 weeks or equal to or greater than 34 weeks or UTD are excluded from the measure.

S.12. **Stratification Details/Variables** *(All information required to stratify the measure results including the stratification variables, definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format with at S.2b)*

Not applicable, the measure is not stratified.

S.13. **Risk Adjustment Type** (Select type. Provide specifications for risk stratification in S.12 and for statistical model in S.14-15)

No risk adjustment or risk stratification

If other:

S.14. **Identify the statistical risk model method and variables** *(Name the statistical method - e.g., logistic regression and list all the risk factor variables. Note - risk model development and testing should be addressed with measure testing under Scientific Acceptability)*

Not Applicable

S.15. **Detailed risk model specifications** *(must be in attached data dictionary/code list Excel or csv file. Also indicate if available at measure-specific URL identified in S.1.)*

Note: Risk model details (including coefficients, equations, codes with descriptors, definitions), should be provided on a separate worksheet in the suggested format in the Excel or csv file with data dictionary/code lists at S.2b.

S.15a. **Detailed risk model specifications** (*if not provided in excel or csv file at S.2b*)

Not Applicable

S.16. **Type of score:**

Rate/proportion

If other:

S.17. **Interpretation of Score** *(Classifies interpretation of score according to whether better quality is associated with a higher score, a lower score, a score falling within a defined interval, or a passing score)*

Better quality = Higher score

S.18. **Calculation Algorithm/Measure Logic** *(Describe the calculation of the measure score as an ordered sequence of steps including identifying the target population; exclusions; cases meeting the target process, condition, event, or outcome; aggregating data; risk adjustment; etc.)*

1. Start processing. Run cases that are included in the PC-Mother Initial Patient Population and pass the edits defined in the Transmission Data Processing Flow: Clinical through this measure.

2. Check ICD-10-CM Principal or Other Diagnosis Codes

a. If at least one of the ICD-10-CM Principal or Other Diagnosis Codes is on Table 11.09.1, the case will proceed to a Measure Category Assignment of B and will not be in the measure population. Stop processing.

b. If none of the ICD-10-CM Principal or Other Diagnosis Codes is on Table 11.09.1, continue processing and proceed to Clinical Trial.

3. Check Clinical Trial

a. If Clinical Trial is missing, the case will proceed to a Measure Category Assignment of X and will be rejected. Stop processing.

b. If Clinical Trial equals Yes, the case will proceed to a Measure Category Assignment of B and will not be in the measure population. Stop processing.
c. If Clinical Trial equals No, continue processing and proceed to Gestational Age.

4. Check Gestational Age
a. If Gestational Age is missing, the case will proceed to a Measure Category Assignment of X and will be rejected. Stop processing.
b. If Gestational Age is less than 24 or greater than or equal to 34 or equal to a Not Unable to Determine Value, the case will proceed to a Measure Category Assignment of B and will not be in the measure population. Stop processing.
c. If Gestational Age is greater than or equal to 24 and less than 34, continue processing and proceed to Antenatal Steroids Initiated.

5. Check Antenatal Steroids Initiated
a. If Antenatal Steroids Initiated is missing, the case will proceed to a Measure Category Assignment of X and will be rejected. Stop processing.
b. If Antenatal Steroids Initiated equals Yes, the case will proceed to a Measure Category Assignment of E and will be in the Numerator Population. Stop processing.
c. If Antenatal Steroids Initiated equals No, continue processing and proceed to Reason for Not Initiating Antenatal Steroids.

6. Check Reason for Not Initiating Antenatal Steroids
a. If Reason for Not Initiating Antenatal Steroids is missing, the case will proceed to a Measure Category Assignment of X and will be rejected. Stop processing.
b. If Reason for Initiating Antenatal Steroids equals Yes, the case will proceed to a Measure Category Assignment of B and will not be in the measure population. Stop processing.
c. If Reason for Not Initiating Antenatal Steroids equals No, the case will proceed to a Measure Category Assignment of D and will be in the measure population. Stop processing.

S.19. Calculation Algorithm/Measure Logic Diagram URL or Attachment *(You also may provide a diagram of the Calculation Algorithm/Measure Logic described above at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
Available at measure-specific web page URL identified in S.1

S.20. Sampling *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*
IF a PRO-PM, identify whether (and how) proxy responses are allowed.
The initial patient population includes patients admitted to the hospital for inpatient acute care are included if they have: ICD-10-PCS Principal or Other Procedure Code as defined in Appendix A, Table 11.01.1, a Patient Age (Admission Date – Birthdate) >= 8 years and < 65 and a Length of Stay (Discharge Date - Admission Date) = 120 days. The sample is taken randomly as follows for a monthly sample:
• Average monthly Initial Patient Population >= 501 results in a minimum random sample size of 101.
• Average monthly Initial Patient Population 126 – 500 results in a minimum random sample size of 20% of the population size.
• Average monthly Initial Patient Population 25 – 125 results in a minimum random sample size of 25.
• Average monthly Initial Patient Population < 25 results in no sampling; 100% Initial Patient Population required

S.21. Survey/Patient-reported data *(If measure is based on a survey, provide instructions for conducting the survey and guidance on minimum response rate.)*
IF a PRO-PM, specify calculation of response rates to be reported with performance measure results.
Not Applicable

S.22. Missing data (specify how missing data are handled, e.g., imputation, delete case.)
Required for Composites and PRO-PMs.
Any file with missing data will result in a measure category assignment of X and rejection of the file from the warehouse, unless the data are not used to process the measure. If the data are used to process the measure and have been reported by the abstractor as No/UTD, the case will result in a measure category assignment of D (i.e., failed measure, not rejected).

S.23. Data Source *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*
If other, please describe in S.24.
Paper Medical Records

S.24. Data Source or Collection Instrument *(Identify the specific data source/data collection instrument e.g. name of database,*

clinical registry, collection instrument, etc.)
IF a PRO-PM, identify the specific PROM(s); and standard methods, modes, and languages of administration.
Each data element in the data dictionary includes suggested data sources. The data are collected using contracted Performance Measurement Systems (vendors) that develop data collection tools based on the measure specifications. The tools are verified and tested by Joint Commission staff to confirm the accuracy and conformance of the data collection tool with the measure specifications. The vendor may not offer the measure set to hospitals until verification has been passed.

S.25. **Data Source or Collection Instrument** *(available at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
No data collection instrument provided

S.26. **Level of Analysis** *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*
Facility, Population : National

S.27. **Care Setting** *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*
Hospital/Acute Care Facility
If other:

S.28. **COMPOSITE Performance Measure** - Additional Specifications *(Use this section as needed for aggregation and weighting rules, or calculation of individual performance measures if not individually endorsed.)*
Not Applicable

2a. Reliability – See attached Measure Testing Submission Form
2b. Validity – See attached Measure Testing Submission Form
PC-03_0476_MeasureTesting_MSF5.0_Data-635787040707209660.doc

NQF #: 0476 **NQF Project:** Perinatal and Reproductive Health Project

2. RELIABILITY & VALIDITY - SCIENTIFIC ACCEPTABILITY OF MEASURE PROPERTIES

Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. **(evaluation criteria)**
Measure testing must demonstrate adequate reliability and validity in order to be recommended for endorsement. Testing may be conducted for data elements and/or the computed measure score. Testing information and results should be entered in the appropriate field. Supplemental materials may be referenced or attached in item 2.1. See guidance on measure testing.

2a2. Reliability Testing. (*Reliability testing was conducted with appropriate method, scope, and adequate demonstration of reliability.*)

2a2.1 Data/Sample *(Description of the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)*:
The PC measure set has been in national use since the 2nd quarter of 2010. It is a requirement of participation in the ORYX initiative that data on all measures in the set are collected. (ORYX is the term used by The Joint Commission to describe the component of the hospital accreditation program which requires data collection and reporting on standardized national performance measures.)
Demographics of organizations collecting and reporting data on these measures are as follows:
163 health care organizations representing various types, locations and sizes:
10 For Profit, 91 Not for Profit, 46 Military Facilities, 9 County, 2 State, 5 Other
15 >=500 beds; 29 250-499 beds; 50 100-249 beds; 69 <100 beds
Located in: AE, AK, AL, AP, AR, AZ, CA, DO, DC, FL, GA, HI, ID, IL, IN, KS, KY, LA, MA, MD, MI, MN, MO, MS, MT, NC, NE, NV, NY, OH,OK, PA, PR, RI, SC, TN, TX, VA, WA, WI, WV
26 performance measurement systems

2a2.2 Analytic Method *(Describe method of reliability testing & rationale)*:
This measure was adapted from NQF-endorsed measure 0476 Appropriate Use of Antenatal Steroids. As such, initial data reliability would have been addressed during the original endorsement. The Joint Commission will be conducting additional reliability studies on this measure as well as the entire PC measure set beginning October 2011.

Currently, hospitals are supported in their data collection and reporting efforts by 26 contracted performance measurement system (PMS) vendors. It is a contractual requirement of Joint Commission listed vendors that the quality and reliability of data submitted to them by contracted health care organizations must be monitored on a quarterly basis. In addition, The Joint Commission analyzes these data by running 17 quality tests on the data submitted into ORYX. (ORYX is the term used by The Joint Commission to describe the component of the hospital accreditation program which requires data collection and reporting on standardized national performance measures). The following is a list of the major tests done on the submitted ORYX data, taken from the 2011 ORYX Performance Measurement System Requirements manual.

- Transmission of complete data
- Usage of individual core measure data received: To understand if the HCO provides the relevant service to treat the relevant population
- Investigation of aberrant data points
- Verification of patient population and sample size
- Identification of missing data elements
- Validation of the accuracy of target outliers
- Data integrity
- Data corrections

Data Element Agreement Rate:
Inter-rater reliability testing methodology utilized by contracted performance measure system vendors as outlined in the contract is as

follows:

- All clinical data elements and all editable demographic elements are scored.
- All measure data are reabstracted with originally abstracted data having been blinded so that the reabstraction is not biased.
- Reabstracted data are compared with originally abstracted data on a data element by data element basis. A data element agreement rate is calculated. Clinical and demographic data are scored separately, and an overall agreement rate is computed.

2a2.3 Testing Results *(Reliability statistics, assessment of adequacy in the context of norms for the test conducted)*:
Data element agreement rates were reported to The Joint Commission for 1Q11. This reflects the findings of 108 hospitals, comprising 13,279 records (100% sample). The following table delineates calculated agreement rates for individual data elements that are used to compute measure rates for PC-03.

Data Elements with a Mismatch -	Mother total n	total d	rate
Antenatal Steroid Administered	20	21	99.16%

This agreement rate is considered to be well within acceptable levels.

2b. VALIDITY. Validity, Testing, including all Threats to Validity: H☐ M☐ L☐ I ☐

2b1.1 Describe how the measure specifications *(measure focus, target population, and exclusions)* **are consistent with the evidence cited in support of the measure focus (***criterion 1c)* **and identify any differences from the evidence:**
This measure focuses on the rate of antenatal steroids administered prior to a live preterm delivery at 24 0/7 to 32 0/7 weeks gestation. The literature supports the focus on patients delivering live newborns within this gestational age range. Also, consonant with the literature, this measure excludes patients with fetal demise. Also excluded from the measure are patients with a length of stay greater than 120 days, and those enrolled in a clinical trial. These exclusions are not addressed in the literature, but are included for this measure in order to harmonize with other CMS/Joint Commission aligned measures.

2b2. Validity Testing. (*Validity testing was conducted with appropriate method, scope, and adequate demonstration of validity.)*

2b2.1 Data/Sample *(Description of the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)*:
As previously mentioned the PC measure set has been in national use since the 2nd quarter of 2010. Demographics of organizations collecting and reporting data on these measures are as follows:
163 health care organizations representing various types, locations and sizes:
10 For Profit, 91 Not for Profit, 46 Military Facilities, 9 County, 2 State, 5 Other
15 >=500 beds; 29 250-499 beds; 50 100-249 beds; 69 <100 beds
Located in: AE, AK, AL, AP, AR, AZ, CA, DO, DC, FL, GA, HI, ID, IL, IN, KS, KY, LA, MA, MD, MI, MN, MO, MS, MT, NC, NE, NV, NY, OH,OK, PA, PR, RI, SC, TN, TX, VA, WA, WI, WV
26 performance measurement systems

2b2.2 Analytic Method *(Describe method of validity testing and rationale; if face validity, describe systematic assessment)*:
Since the measure has been in national use, continued face validity of the measure has been determined through analysis of feedback from measure users. The Joint Commission provides a web-based application with which measure users can provide feedback regarding appropriateness of measure specifications, request clarification of specifications, and/or provide other comments pertinent to the measure. This feedback is systematically continually reviewed in order to identify trends and to identify areas of the measure specifications that require clarification or revision. Additionally, Joint Commission staff continually monitors the national literature and environment in order to assess continued validity of this measure.

As noted previously, The Joint Commission is currently performing reliability site visits. A component of these visits will include focus group interviews with hospital staff working with the PC measures to obtain feedback regarding the validity of the measures and suggestions for further refinement of the specifications.

2b2.3 Testing Results *(Statistical results, assessment of adequacy in the context of norms for the test conducted; if face validity, describe results of systematic assessment)*:
Analysis of feedback obtained via our automated feedback system reveals slightly more than 90 submissions regarding specifications for this measure since its implementation in 2010. Predominant themes of these submissions involved questions regarding clarification of the data elements Antenatal Steroid Administered and Reason for Not Administering Antenatal Steroid with respect to timing of the repeat dose and implied reasons for not administering a full course of antenatal steroids. Additional notes for abstractors were added to

the data elements for clarification. In addition, the denominator excluded population and algorithm were revised to exclude patients with fetal demise with an additional ICD-9-CM diagnosis code table. The gestational age range for the denominator statement was also revised to exclude patients with a gestational age of 32 1/7 to 32 6/7 weeks of gestation, since the data element for gestational age instructs the abstractors to round gestational age down to the nearest completed week.

POTENTIAL THREATS TO VALIDITY. (*All potential threats to validity were appropriately tested with adequate results.*)

2b3. Measure Exclusions. (*Exclusions were supported by the clinical evidence in 1c or appropriately tested with results demonstrating the need to specify them.*)

2b3.1 Data/Sample for analysis of exclusions *(Description of the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)***:**
As previously mentioned the PC measure set has been in national use since the 2nd quarter of 2010. Demographics of organizations collecting and reporting data on these measures are as follows:
163 health care organizations representing various types, locations and sizes:
10 For Profit, 91 Not for Profit, 46 Military Facilities, 9 County, 2 State, 5 Other
15 >=500 beds; 29 250-499 beds; 50 100-249 beds; 69 <100 beds
Located in: AE, AK, AL, AP, AR, AZ, CA, DO, DC, FL, GA, HI, ID, IL, IN, KS, KY, LA, MA, MD, MI, MN, MO, MS, MT, NC, NE, NV, NY, OH,OK, PA, PR, RI, SC, TN, TX, VA, WA, WI, WV
26 performance measurement systems

2b3.2 Analytic Method (*Describe type of analysis and rationale for examining exclusions, including exclusion related to patient preference)***:**
Measure exclusions that were not derived directly from the evidence are presented below. Please note that these are population exclusions that are necessary to ensure consistency in all measures in this 5 measure set.
These exclusions were analyzed for frequency of occurrence. An issue that is of great concern to users of this measure is that due to the presence of exceptions to the measure, attainment of a 100% measure rate is not possible. Because of the role of this measure in the current Joint Commission accreditation process, this is especially troubling to measure users. This concern is the basis for a number of the non-evidence-based exclusions to these measures. Additional reasons for these population exclusions are enumerated in our response to section 2b1.1 above. The following measure exclusions that were not derived directly from the evidence are as follows:

1. Patients with LOS <120 days
2. Patients less than 8 years of age or greater than or equal to 65 years of age
3. Patients enrolled in clinical trials

2b3.3 Results *(Provide statistical results for analysis of exclusions, e.g., frequency, variability, sensitivity analyses):*
N=353,671
1. Patients who have a length of stay (LOS) greater than 120 days =0%
2. Patients less than 8 years of age or greater than or equal to 65 years of age=0%
3. Patients enrolled in clinical trials =0.06%

2b4. Risk Adjustment Strategy. (*For outcome measures, adjustment for differences in case mix (severity) across measured entities was appropriately tested with adequate results.*)

2b4.1 Data/Sample *(Description of the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)***:**
Not Applicable

2b4.2 Analytic Method (*Describe methods and rationale for development and testing of risk model or risk stratification including selection of factors/variables)***:**
Not Applicable

2b4.3 Testing Results *(Statistical risk model: Provide quantitative assessment of relative contribution of model risk factors; risk model performance metrics including cross-validation discrimination and calibration statistics, calibration curve and risk decile plot, and assessment of adequacy in the context of norms for risk models. Risk stratification: Provide quantitative assessment of relationship of risk factors to the outcome and differences in outcomes among the strata)***:**
Not Applicable

2b4.4 If outcome or resource use measure is not risk adjusted, provide rationale and analyses to justify lack of adjustment:
Not Applicable

2b5. Identification of Meaningful Differences in Performance. (*The performance measure scores were appropriately analyzed and discriminated meaningful differences in quality.*)

2b5.1 Data/Sample *(Describe the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)***:**
As previously mentioned the PC measure set has been in national use since the 2nd quarter of 2010. Demographics of organizations collecting and reporting data on these measures are as follows:
163 health care organizations representing various types, locations and sizes:
10 For Profit, 91 Not for Profit, 46 Military Facilities, 9 County, 2 State, 5 Other
15 >=500 beds; 29 250-499 beds; 50 100-249 beds; 69 <100 beds
Located in: AE, AK, AL, AP, AR, AZ, CA, DO, DC, FL, GA, HI, ID, IL, IN, KS, KY, LA, MA, MD, MI, MN, MO, MS, MT, NC, NE, NV, NY, OH,OK, PA, PR, RI, SC, TN, TX, VA, WA, WI, WV
26 performance measurement systems

2b5.2 Analytic Method *(Describe methods and rationale to identify statistically significant and practically/meaningfully differences in performance)***:**
The method used to analyze meaningful differences in performance at The Joint Commission is Target Analysis. The object of target analysis is to compare a health care organizations (HCO) data against a comparative norm for the purpose of evaluating performance improvement opportunities. When an organization's performance level is statistically significantly different from a comparative norm, it is considered a statistical deviation. A statistical deviation may be desirable or undesirable depending on the "direction of improvement" of the measure.
There are two components to the target analysis methodology used at The Joint Commission. Given the national average for a performance measure, a target range is constructed. Using generalized linear mixed models methodology (also known as hierarchical models), a predicted estimate of an HCO's performance, with a corresponding 95% confidence interval, is generated. This confidence interval is compared to the target range, to determine the HCOs' rating. The estimate of the organization's true performance is based on both the data from that organization and on data from the entire set of reporting organizations.

2b5.3 Results *(Provide measure performance results/scores, e.g., distribution by quartile, mean, median, SD, etc.; identification of statistically significant and meaningfully differences in performance)*:
PC-03 Distribution of Outliers

2011 1st Quarter Data:
Scores on this measure: N=79, Mean 64.6%, SD 0.40336
10th Percentile= 0%
25th Percentile= 25%
50th Percentile= 80%
75th Percentile= 100%
90th Percentile= 100%

79 (100%) Neutral – results not significantly different from target range

2b6. Comparability of Multiple Data Sources/Methods. (*If specified for more than one data source, the various approaches result in comparable scores.*)

2b6.1 Data/Sample *(Describe the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)***:**
Multiple data sources are not used for this measure.

2b6.2 Analytic Method *(Describe methods and rationale for testing comparability of scores produced by the different data sources specified in the measure)***:**
Not applicable

2b6.3 Testing Results *(Provide statistical results, e.g., correlation statistics, comparison of rankings; assessment of adequacy in the*

context of norms for the test conducted):
Not applicable

2c. Disparities in Care: H☐ M☐ L☐ I ☐ NA☐ (*If applicable, the measure specifications allow identification of disparities.*)

2c.1 If measure is stratified for disparities, provide stratified results *(Scores by stratified categories/cohorts)*: The measure is not stratified for disparities.

2c.2 If disparities have been reported/identified (e.g., in 1b), but measure is not specified to detect disparities, please explain: The California Perinatal Quality Care Collaborative database for California births reports disparities for Hispanic mothers, mothers younger than age 20, mothers without prenatal care, mothers giving birth vaginally and mothers with a diagnosis of fetal distress. This measure is not stratified. The Joint Commission does not currently capture date elements for race or ethnicity because these data elements have not been shown to be reliably collectable due to the fact that no national standardized definitions exist for these data elements. Also, not all hospitals collect race and ethnicity. In the future, it may be feasible for The Joint Commission to explore how race and ethnicity and other relevant disparity data, might be collected reliably. Future measure data could also be evaluated according to sex, age, presence of prenatal care, type of delivery and geographic location.

2.1-2.3 Supplemental Testing Methodology Information:

Steering Committee: Overall, was the criterion, *Scientific Acceptability of Measure Properties*, met?
(*Reliability and Validity must be rated moderate or high*) Yes☐ No☐
Provide rationale based on specific subcriteria:

If the Committee votes No, STOP

NEW (since 2012 endorsement): Data for Empirical Testing

2b2. VALIDITY TESTING
2b2.1. What level of validity testing was conducted? (*may be one or both levels*)

☐ **Critical data elements** (*data element validity must address ALL critical data elements*)
☐ **Performance measure score**
☒ **Empirical validity testing**
☐ **Systematic assessment of face validity of performance measure score as an indicator** of quality or resource use (*i.e., is an accurate reflection of performance on quality or resource use and can distinguish good from poor performance*)

2b2.2. For each level of testing checked above, describe the method of validity testing and what it tests (*describe the steps—do not just name a method; what was tested, e.g., accuracy of data elements compared to authoritative source, relationship to another measure as expected; what statistical analysis was used)*

ICD-9 to ICD-10 Conversion Process:

The goal was to convert ICD-9 to ICD-10 equivalent codes, consistent with the clinical intent of the original measure specifications. The Joint Commission worked with a certified coding expert throughout the conversion process. The 3M Coding Conversion Tool was utilized, including forward mapping of ICD-9 codes to ICD-10 codes as well as reverse mapping from ICD-10 to ICD-9 to ensure appropriateness. MSDRGs and instructions in the tabular index were also examined to ensure appropriate code mapping. Crosswalks comprising ICD-9 codes mapped to ICD-10 codes were created and reviewed by members of the Technical Advisory Panel, CMS subcontractors, and performance measurement system vendors prior to being posted for a 12 month public comment period. Feedback from the field indicated that the crosswalks generally were mapped correctly. Minor modifications to the code tables were made as needed. Final code tables were published in early 2015, well in advance of the mandated date of October 1, 2015.

Perinatal Care (PC) Initial Patient Population

The PC measure set is unique in that there are two distinct Initial Patient Populations within the measure set, mothers (PC-01, PC-02, PC-03) and newborns. (PC-04, PC-05).

Subpopulation Mothers
group if they have: ICD-9-CM Principal or Other Diagnosis Code as defined in Appendix A, Tables 11.01, 11.02, 11.03, or 11.04, a Patient Age (Admission Date Birthdate) >= 8 years and < 65 and a Length of Stay (Discharge Date - Admission Date) ≤ 120 days. PC-03- Antenatal Steroids belongs to the above population.

The data used to measure the validity of the PC measure are comprised of data from the third and fourth quarters of 2014, and the first and second quarters of 2015. 1,345 hospitals submitted 2,695,467 inpatient records for all the elected PC measures. The hospitals included in the analysis reported one year of data and had 30 or more denominator cases in the analysis period.
Measure convergent validity for PC-03 was assessed using hospitals patient level data from The Joint Commission warehouse. Measure specifications, including population identification, numerator and denominator statements, exclusions, and data elements and their definitions were found to be understandable, retrievable, and relevant in previous validity testing.

2b2.3. What were the statistical results from validity testing? (*e.g., correlation; t-test*)

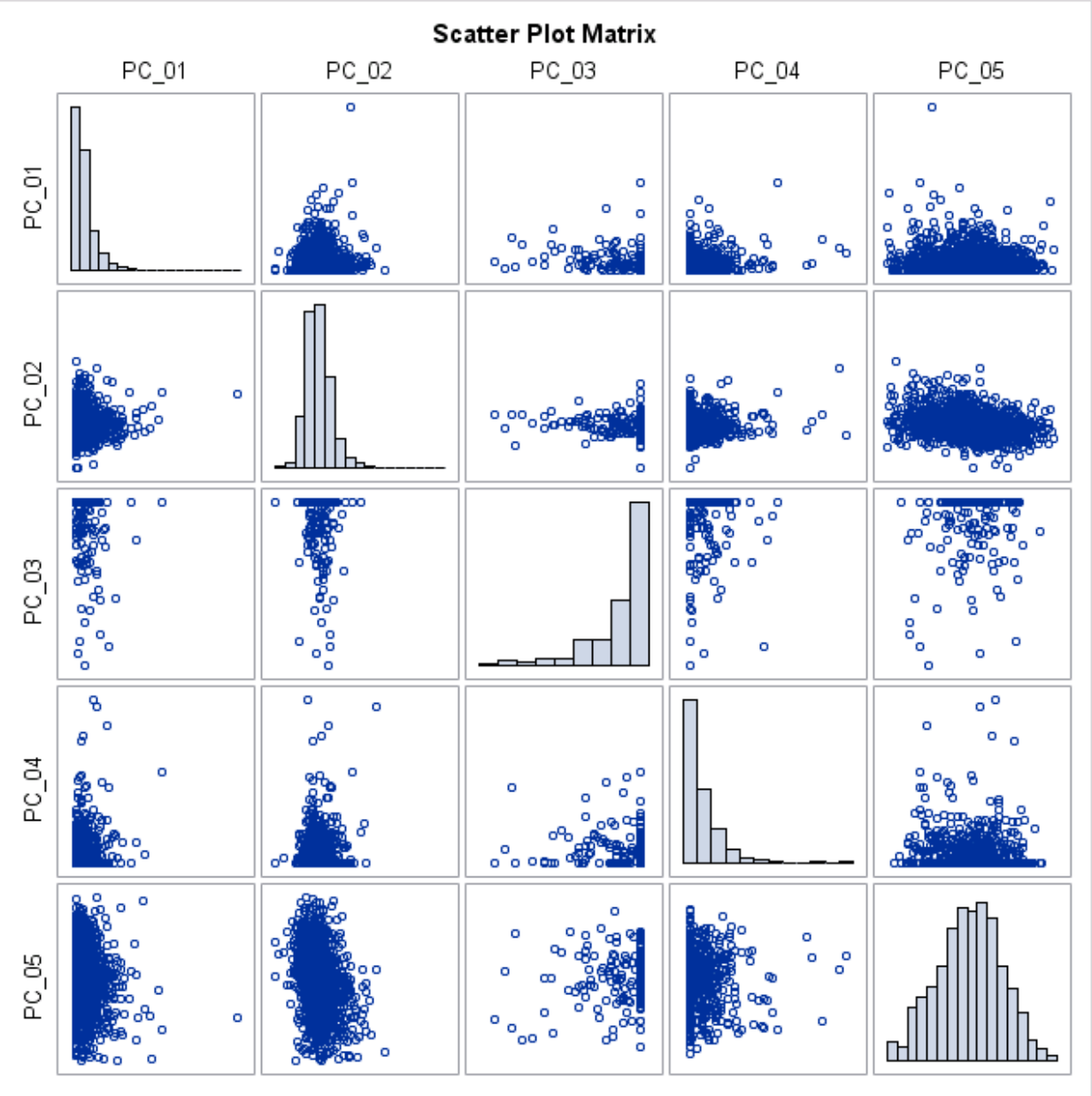
Overall descriptive statistics for sub population MOTHER:
N=1,345 hospitals
n = 2,695,467 records submitted

Descriptive statistics for PC-03 measure
N = 162 hospitals
Number of records = 390,903
Mean = 98%
Min = 85%
Percentile 10%: 94%
Percentile 25%: 97%
Median: 99%
Percentile 75%: 100%
Percentile 90%: 100%
Max = 100%

Simple Statistics

Variable	N	Mean	Std Dev	Median	Minimum	Maximum
PC_01	1237	0.02753	0.03803	0.01734	0	0.51240
PC_02	1345	0.26287	0.07974	0.25410	0	1.00000
PC_03	162	0.97762	0.03311	0.99425	0.84615	1.00000
PC_04	523	0.05267	0.08432	0.02203	0	0.66129
PC_05	1352	0.49198	0.19284	0.50190	0.00317	1.00000

Spearman Correlation Coefficients Prob > |r| under H0: Rho=0 Number of Observations					
	PC_01	PC_02	PC_03	PC_04	PC_05
PC_01	1.00000 1237	0.06843 0.0163 1231	-0.26960 0.0006 159	0.10724 0.0169 496	-0.03538 0.2137 1237
PC_02	0.06843 0.0163 1231	1.00000 1345	-0.18318 0.0196 162	0.02807 0.5218 523	-0.32009 <.0001 1343
PC_03	-0.26960 0.0006 159	-0.18318 0.0196 162	1.00000 162	-0.03117 0.7030 152	0.07729 0.3283 162
PC_04	0.10724 0.0169 496	0.02807 0.5218 523	-0.03117 0.7030 152	1.00000 523	-0.03560 0.4165 523
PC_05	-0.03538 0.2137 1237	-0.32009 <.0001 1343	0.07729 0.3283 162	-0.03560 0.4165 523	1.00000 1352



The Spearman rank-order correlation is a nonparametric measure of association based on the ranks of the data values by measure PC-03 and hospitals. We used this methodology because of the skewness of the distribution of the measure rates.

.

2b2.4. What is your interpretation of the results in terms of demonstrating validity? (i.*e., what do the results mean and what are the norms for the test conducted?*)

The correlation of PC-03 with the other PC measures in the PC measure set indicates that the correlations with two other PC measures are moderate and statistically significant. Although 90% of the hospital measure rates fall between 94 and 100%, there are some hospitals whose performance on this measure is significantly below the national average.

2b3. EXCLUSIONS ANALYSIS .

2b3.1. Describe the method of testing exclusions and what it tests (*describe the steps—do not just name a method; what was tested, e.g., whether exclusions affect overall performance scores; what statistical analysis was used*)

There were 390,903 admissions selected from the initial cohort. From among the 390,903 admissions in 162 hospitals, the descriptive statistics are given below.

The following exclusions were analyzed by subpopulation and measure for frequency and variability across providers:

- Less than 8 years of age
- Greater than or equal to 65 years of age
- Length of Stay >120 days
- Enrolled in clinical trials
- Documented *Reason for Not Initiating Antenatal Steroid Therapy*
- *ICD-9-CM Principal Diagnosis Code or ICD-9-CM Other Diagnosis Codes* for fetal demise as defined in Appendix A, Table 11.09.1
- *Gestational Age* < 24 or >= 34 weeks or UTD

2b3.2. What were the statistical results from testing exclusions? (*include overall number and percentage of individuals excluded, frequency distribution of exclusions across measured entities, and impact on performance measure scores*)

Exclusion Subpopulation 1 – PC-03:

Less than 8 years of age
Exclusion: Included in the initial population exclusion

Greater than or equal to 65 years of age
Exclusion*:* Included in the initial population exclusion

Length of Stay >120 days
Exclusion: Included in the initial population exclusion

Exclusion: Patients enrolled in clinical trials
Overall Occurrence n = 271
Overall Occurrence Percentage: 0.07%
Minimum: 0%
10th Percentile: 0%
Median: 0%
90th Percentile: 0.15%
Maximum: 6.59%

Exclusion: Documented Reason for Not Initiating Antenatal Steroid Therapy
Overall Occurrence n = 5,272
Overall Occurrence Percentage: 1.35%
Minimum: 0%
10th Percentile: 0.045%
Median: 0.20%
90th Percentile: 0.54%
Maximum: 71.17%

Exclusion: ICD-9-CM Principal Diagnosis Code or ICD-9-CM Other Diagnosis Codes for fetal demise as defined in Appendix A, Table 11.09.1
Overall Occurrence n = 36
Overall Occurrence Percentage: 0.009%
Minimum: 0%

10th Percentile: 0%
Median 0%
90th Percentile: 0.034%
Maximum: 0.16%

Exclusion: Gestational Age < 24 or >= 34 weeks or UTD
Overall Occurrence n = 376,377
Overall Occurrence Percentage: 96%
Minimum: 85%
10th Percentile: 92%
Median: 96%
90th Percentile: 97%
Maximum: 98%

2b3.3. What is your interpretation of the results in terms of demonstrating that exclusions are needed to prevent unfair distortion of performance results? (*i.e., the value outweighs the burden of increased data collection and analysis. Note: **If patient preference is an exclusion**, the measure must be specified so that the effect on the performance score is transparent, e.g., scores with and without exclusion*)

The overall frequency of most exclusions is low for those in the measure denominator. The only exclusion that has an overall Occurrence percentage equal 96% is "Gestational Age < 24 or >= 34 weeks or UTD". The high percentage is justified by the scope of measure PC-03. The difference between the 10th and 90th percentiles of the distribution of exclusion rates is narrow indicating that the occurrence is random and likely would not bias performance results.

It is believed that all of the exclusions should be retained for the following reasons:

Exclusion: Patients who have a *Length of stay* greater than 120 days
Rationale: Included for this measure in order to harmonize with other CMS/Joint Commission aligned measures.

Exclusion: Patients enrolled in a Clinical Trial
Rationale: Only capture patients not enrolled in clinical trials studying pregnant patients or newborns.

Exclusion Documented Reason for Not Initiating Antenatal Steroid Therapy
Rationale: Reasons for not initiating antenatal steroids may include fetal distress, imminent delivery or other reasons documented by physician/advanced practice nurse (APN)/physician assistant (PA)/certified nurse midwife (CNM).

Exclusion: ICD-9-CM Principal Diagnosis Code or ICD-9-CM Other Diagnosis Codes for fetal demise as defined in Appendix A, Table 11.09.1
Rationale: Table 11.09.1 contains diagnosis codes for fetal demise in order to exclude the case from the denominator, since there is no reason to initiate antenatal steroids prior to delivery.

Exclusion: Gestational Age < 24 or >= 34 weeks or UTD
Rationale: The denominator population is limited to patients ≥ 24 to < 34 weeks of completed gestation. Patients with UTD for gestational age typically have had no prenatal care.

3. Feasibility

Extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

3a. Byproduct of Care Processes

For clinical measures, the required data elements are routinely generated and used during care delivery (e.g., blood pressure, lab test, diagnosis, medication order).

3a.1. **Data Elements Generated as Byproduct of Care Processes.**

generated by and used by healthcare personnel during the provision of care, e.g., blood pressure, lab value, medical condition, Coded by someone other than person obtaining original information (e.g., DRG, ICD-9 codes on claims), Abstracted from a record by someone other than person obtaining original information (e.g., chart abstraction for quality measure or registry)

If other:

3b. Electronic Sources

The required data elements are available in electronic health records or other electronic sources. If the required data are not in electronic health records or existing electronic sources, a credible, near-term path to electronic collection is specified.

3b.1. **To what extent are the specified data elements available electronically in defined fields?** *(i.e., data elements that are needed to compute the performance measure score are in defined, computer-readable fields)*

Some data elements are in defined fields in electronic sources

3b.2. **If ALL the data elements needed to compute the performance measure score are not from electronic sources, specify a credible, near-term path to electronic capture, OR provide a rationale for using other than electronic sources.**

PC-03 is in the queue to be re-engineered as an eCQM as resources permit in the future.

3b.3. **If this is an eMeasure, provide a summary of the feasibility assessment in an attached file or make available at a measure-specific URL.**

Attachment:

3c. Data Collection Strategy

Demonstration that the data collection strategy (e.g., source, timing, frequency, sampling, patient confidentiality, costs associated with fees/licensing of proprietary measures) can be implemented (e.g., already in operational use, or testing demonstrates that it is ready to put into operational use). For eMeasures, a feasibility assessment addresses the data elements and measure logic and demonstrates the eMeasure can be implemented or feasibility concerns can be adequately addressed.

3c.1. **Describe what you have learned/modified as a result of testing and/or operational use of the measure regarding data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, time and cost of data collection, other feasibility/implementation issues.**

IF a PRO-PM, consider implications for both individuals providing PROM data (patients, service recipients, respondents) and those whose performance is being measured.

At the present time, hospitals using this performance measure generally collect measure data via manual review of the paper medical record, the EMR or a combination of both. Collected data are submitted to The Joint Commission on a quarterly basis, by way of contracted performance measurement system vendors, as described previously. Specifications for this measure are freely available to anyone who wishes to use the measure. Feedback from hospitals using this measure indicates that required data elements are generally available in the medical record, and measure specifications are robust and easy to understand. As described above, as feedback from measure users has indicated the need for clarification or revision of measure specifications, this has taken place.

3c.2. **Describe any fees, licensing, or other requirements to use any aspect of the measure as specified** *(e.g., value/code set, risk model, programming code, algorithm)*.

There are no fees or licensing requirements to use the Joint Commission performance measures, all of which are in the public domain.

4. Usability and Use

Extent to which potential audiences (e.g., consumers, purchasers, providers, policy makers) are using or could use performance results for both accountability and performance improvement to achieve the goal of high-quality, efficient healthcare for individuals or populations.

4a. Accountability and Transparency

Performance results are used in at least one accountability application within three years after initial endorsement and are publicly reported within six years after initial endorsement (or the data on performance results are available). If not in use at the time of initial endorsement, then a credible plan for implementation within the specified timeframes is provided.

4.1. Current and Planned Use

NQF-endorsed measures are expected to be used in at least one accountability application within 3 years and publicly reported within 6 years of initial endorsement in addition to performance improvement.

Planned	Current Use (for current use provide URL)
Quality Improvement with Benchmarking (external benchmarking to multiple organizations)	Public Reporting Quality Check® http://www.qualitycheck.org/consumer/searchQCR.aspx Regulatory and Accreditation Programs Hospital Accreditation Program http://jointcommission.org Quality Improvement (Internal to the specific organization) Perinatal Care Certification http://www.jointcommission.org/certification/perinatal_care_certification.aspx

4a.1. For each CURRENT use, checked above, provide:

- Name of program and sponsor
- Purpose
- Geographic area and number and percentage of accountable entities and patients included

- Name of program and sponsor: Quality Check®; The Joint Commission
- Purpose: A public website that allows consumers to: search for accredited and certified organizations by city and state, by name or by zip code (up to 250 miles); find organizations by type of service provided within a geographic area; download free hospital performance measure results; and, print a list of Joint Commission certified disease-specific care programs and health care staffing firms.
- Geographic area and number and percentage of accountable entities and patients included: Nationwide; 3300 Joint Commission-accredited hospitals (2014)

- Name of program and sponsor Hospital Accreditation Program; The Joint Commission
- Purpose: An accreditation program that recognizes hospitals that meet standard requirements to provide safe and effective patient care.
- Geographic area and number and percentage of accountable entities and patients included Nationwide; 3300 Joint Commission-accredited hospitals (2014)

- Name of program and sponsor America's Hospitals: Improving Quality and Safety – The Joint Commission's Annual Report ; The Joint Commission
- Purpose: The annual report summarizes the performance of Joint Commission-accredited hospitals on 46 accountability measures of evidence-based care processes closely linked to positive patient outcomes, and provides benchmarks from Top Performer on Key Quality Measures® hospitals.
- Geographic area and number and percentage of accountable entities and patients included Nationwide; 3300 Joint Commission-accredited hospitals (2014)

- Name of program and sponsor Perinatal Care Certification; The Joint Commission

- Purpose: A certification program that recognizes hospitals that have achieved integrated, coordinated, patient-centered care for clinically uncomplicated pregnancies and births.
- Geographic area and number and percentage of accountable entities and patients included Nationwide; Twelve Joint Commission-accredited hospitals (2016)

4a.2. **If not currently publicly reported OR used in at least one other accountability application (e.g., payment program, certification, licensing) what are the reasons?** (*e.g., Do policies or actions of the developer/steward or accountable entities restrict access to performance results or impede implementation?*)
Not Applicable

4a.3. **If not currently publicly reported OR used in at least one other accountability application, provide a credible plan for implementation within the expected timeframes -- any accountability application within 3 years and publicly reported within 6 years of initial endorsement.** (*Credible plan includes the specific program, purpose, intended audience, and timeline for implementing the measure within the specified timeframes. A plan for accountability applications addresses mechanisms for data aggregation and reporting.*)
Not Applicable

4b. Improvement

Progress toward achieving the goal of high-quality, efficient healthcare for individuals or populations is demonstrated. If not in use for performance improvement at the time of initial endorsement, then a credible rationale describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.

4b.1. **Progress on Improvement. (Not required for initial endorsement unless available.)**
Performance results on this measure (current and over time) should be provided in 1b.2 and 1b.4. Discuss:

- **Progress (trends in performance results, number and percentage of people receiving high-quality healthcare)**
- **Geographic area and number and percentage of accountable entities and patients included**

The rate of patients receiving antenatal steroids prior to premature deliveries has increased dramatically from 63.3% in 2010 with 114 hospitals reporting to 91.6% in 2014 with 1133 hospitals reporting based on Joint Commission OPRYX performance measurement data from across the country. The 2014 inclusion of all accredited hospitals with 1100 births or more annually now captures essentially all accredited hospitals delivering premature newborns making the data more reflective of overall national performance. The rate of patients receiving this important intervention has continued to improve every year since data collection began in 2010.

4b.2. **If no improvement was demonstrated, what are the reasons? If not in use for performance improvement at the time of initial endorsement, provide a credible rationale that describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.**
Not Applicable

4c. Unintended Consequences

The benefits of the performance measure in facilitating progress toward achieving high-quality, efficient healthcare for individuals or populations outweigh evidence of unintended negative consequences to individuals or populations (if such evidence exists).

4c.1. **Were any unintended negative consequences to individuals or populations identified during testing; OR has evidence of unintended negative consequences to individuals or populations been reported since implementation? If so, identify the negative unintended consequences and describe how benefits outweigh them or actions taken to mitigate them.**
Unintended Consequence:
Cases were inappropriately failing when the repeat dose of antenatal steroids was not given due to the delivery occurring prior to the routinely scheduled repeat dose being ordered at the time the first dose was given.
Mitigating Action:
The data element Antenatal Steroids Administered was changed to Antenatal Steroids Initiated to capture initiation of antenatal steroids instead of a full course.

Unintended Consequence:
Patients who did not receive prenatal care were inappropriately included in the measure denominator, as the gestational age data element was abstracted as unable to be determined (UTD).
Mitigating Action:
In order to avoid penalizing hospitals, cases with UTD were removed from the measure population.

Unintended Consequence:
Some hospitals have reported lower rates due to small denominator populations as a result of sampling.
Mitigating Action:
Vital Records reports, delivery logs and clinical information systems were added as acceptable data sources to help hospitals identify all cases with 24 and less than 34 weeks gestation, so that 100% of these cases could be reviewed to increase the denominator population size.

5. Comparison to Related or Competing Measures

If a measure meets the above criteria and there are endorsed or new related measures (either the same measure focus or the same target population) or competing measures (both the same measure focus and the same target population), the measures are compared to address harmonization and/or selection of the best measure.

5. Relation to Other NQF-endorsed Measures
Are there related measures (conceptually, either same measure focus or target population) or competing measures (conceptually both the same measure focus and same target population)? If yes, list the NQF # and title of all related and/or competing measures.
No

5.1a. **List of related or competing measures (selected from NQF-endorsed measures)**

5.1b. **If related or competing measures are not NQF endorsed please indicate measure title and steward.**

5a. Harmonization

The measure specifications are harmonized with related measures;
OR
The differences in specifications are justified

5a.1. **If this measure conceptually addresses EITHER the same measure focus OR the same target population as NQF-endorsed measure(s):**
Are the measure specifications completely harmonized?

5a.2. **If the measure specifications are not completely harmonized, identify the differences, rationale, and impact on interpretability and data collection burden.**
Not Applicable

5b. Competing Measures

The measure is superior to competing measures (e.g., is a more valid or efficient way to measure);
OR
Multiple measures are justified.

5b.1. **If this measure conceptually addresses both the same measure focus and the same target population as NQF-endorsed measure(s):**
Describe why this measure is superior to competing measures (e.g., a more valid or efficient way to measure quality); OR provide a rationale for the additive value of endorsing an additional measure. (Provide analyses when possible.)
Not Applicable

Appendix

A.1 Supplemental materials may be provided in an appendix. All supplemental materials (such as data collection instrument or methodology reports) should be organized in one file with a table of contents or bookmarks. If material pertains to a specific submission form number, that should be indicated. Requested information should be provided in the submission form and required attachments. There is no guarantee that supplemental materials will be reviewed.

Available at measure-specific web page URL identified in S.1 **Attachment:**

Contact Information

Co.1 **Measure Steward (Intellectual Property Owner):** The Joint Commission
Co.2 **Point of Contact:** Ann, Watt, awatt@jointcommission.org, 630-792-5944-
Co.3 **Measure Developer if different from Measure Steward:** The Joint Commission
Co.4 **Point of Contact:** Ann, Watt, awatt@jointcommission.org, 630-792-5944-

Additional Information

Ad.1 **Workgroup/Expert Panel involved in measure development**

Provide a list of sponsoring organizations and workgroup/panel members' names and organizations. Describe the members' role in measure development.

Michael Ross, MD, MPH (Chair)
Harbor-UCLA Medical Center
Torrance, CA

Wanda Barfield, MD, MPH
Centers for Disease Control and Prevention
Atlanta, GA

Kenneth E. Brown, MD, MBA, FACOG, FACHE
Woman´s Hospital
Lafayette, LA

Martin McCaffrey, MD
UNC North Carolina Children's Hospital
Chapel Hill, NC

Cathy Collins-Fulea, MSN, CNM
Henry Ford Hospital
Detroit, MI

Janet H. Muri, MBA
National Perinatal Information Center/
Quality Analytic Services
Providence, RI

Kathleen Simpson, PhD, RNC, FAAN
St. John's Mercy Medical Center
St. Louis, MO

Michael Socol, MD
Northwestern University Medical School
Chicago, IL

Rebecca Zimmermann, MPP
America's Health Insurance Plans
Washington, DC

The technical advisory panel (TAP) members determined priority areas that could be evaluated to improve care related to perinatal care during the development timeframe. After implementation, minor revisions, acknowledged by TAP representatives, were made to improve clarity. Hospital feedback will be reviewed during the reliability testing phase of the project to assist the TAP in making the final measure recommendations.

Measure Developer/Steward Updates and Ongoing Maintenance

Ad.2 **Year the measure was first released:** 2010
Ad.3 **Month and Year of most recent revision:** 10, 2015
Ad.4 **What is your frequency for review/update of this measure?** Biannual
Ad.5 **When is the next scheduled review/update for this measure?** 02, 2016

Ad.7 **Disclaimers:**

Ad.8 **Additional Information/Comments:**

MEASURE WORKSHEET

This document summarizes the evaluation of the measure as it progresses through NQF's Consensus Development Process (CDP). The information submitted by measure developers/stewards is included after the Brief Measure Information, Preliminary Analysis, and Pre-meeting Public and Member Comments sections.

To navigate the links in the worksheet: Ctrl + click link to go to the link; ALT + LEFT ARROW to return

Brief Measure Information

NQF #: 0478
Measure Title: Neonatal Blood Stream Infection Rate (NQI 03)
Measure Steward: Agency for Healthcare Research and Quality
Brief Description of Measure: Discharges with healthcare-associated blood stream infection per 1,000 discharges for newborns and outborns with birth weight of 500 grams or more but less than 1,500 grams; with gestational age between 24 and 30 weeks; or with birth weight of 1,500 grams or more and death, an operating room procedure, mechanical ventilation, or transferring from another hospital within two days of birth. Excludes discharges with a length of stay less than 3 days and discharges with a principal diagnosis of sepsis, sepsis or bacteremia, or newborn bacteremia.
Developer Rationale: Low birth weight, or premature, and critically ill infants are at increased risk for sepsis or blood stream infections due to immature immune systems, immature skin barriers, and invasive devices such as central venous access or arterial access, ventilation or feeding tubes. Septicemia is one of the most common neonatal infection in neonatal intensive care units. Processes such as hand washing, evidence based vascular access procedures and central line care, and appropriate administration of prophylactic antibiotics can lower rates of neonatal blood stream infection.

Numerator Statement: Discharges, among cases meeting the inclusion and exclusion rules for the denominator, with either:

- any secondary ICD-9-CM or ICD-10 CM diagnosis codes for other septicemia; or
- any secondary ICD-9-CM or ICD-10 CM diagnosis codes for newborn septicemia or bacteremia and
- any secondary ICD-9-CM or ICD-10 CM diagnosis codes for staphylococcal or Gram-negative bacterial infection

Denominator Statement: All newborns and outborns with either:

- a birth weight of 500 to 1,499 grams (Birth Weight Categories 2, 3, 4 and 5); or
- any-listed ICD-9-CM or ICD-10 CM diagnosis codes for gestational age between 24 and 30 weeks; or
- a birth weight greater than or equal to 1,500 grams (Birth Weight Category 6, 7, 8, or 9) and death (DISP=20); or
- a birth weight greater than or equal to 1,500 grams (Birth Weight Category 6, 7, 8, or 9) and any-listed ICD-9-CM or ICD-10 PCS procedure codes for operating room procedure; or
- a birth weight greater than or equal to 1,500 grams (Birth Weight Category 6, 7, 8, or 9) and any-listed ICD-9-CM or ICD-10 PCS procedure codes for mechanical ventilation; or
- a birth weight greater than or equal to 1,500 grams (Birth Weight Category 6, 7, 8, or 9) and transferring from another health care facility within two days of birth

See Pediatric Quality Indicators Appendices:

- Appendix A – Operating Room Procedure Codes
- Appendix I – Definitions of Neonate, Newborn, Normal Newborn, and Outborn
- Appendix L – Low Birth Weight Categories

Denominator Exclusions: Exclude cases:

- with a principal ICD-9-CM or ICD-10-CM diagnosis code (or secondary diagnosis present on admission†) for sepsis
- with a principal ICD-9-CM or ICD-10-CM diagnosis code (or secondary diagnosis present on admission†) for sepsis or bacteremia
- with a principal ICD-9-CM or ICD-10-CM diagnosis code (or secondary diagnosis present on admission†) for staphylococcal or Gram-negative bacterial infection

- with birth weight less than 500 grams (Birth Weight Category 1)
- with length of stay less than 3 days
- with missing gender (SEX=missing), age (AGE=missing), quarter (DQTR=missing), year (YEAR=missing) or principal diagnosis (DX1=missing)

† Only for cases that otherwise qualify for the numerator.

Measure Type: Outcome
Data Source: Administrative claims
Level of Analysis: Facility

IF Endorsement Maintenance – Original Endorsement Date: Oct 24, 2008 **Most Recent Endorsement Date:** Apr 02, 2012

Maintenance of Endorsement -- Preliminary Analysis

To maintain NQF endorsement endorsed measures are evaluated periodically to ensure that the measures still meets the NQF endorsement criteria ("maintenance"). The emphasis for maintaining endorsement is focused on how effective the measure is for promoting improvements in quality. Endorsed measures should have some experience from the field to inform the evaluation. The emphasis for maintaining endorsement is noted for each criterion.

Criteria 1: Importance to Measure and Report

1a. Evidence
Maintenance measures – less emphasis on evidence unless there is new information or change in evidence since the prior evaluation.

1a. Evidence. The evidence requirements for a health outcomes measure include providing rationale that supports the relationship of the health outcome to processes or structures of care. The guidance for evaluating the clinical evidence asks if the relationship between the measured health outcome and at least one clinical action is identified and supported by the stated rationale.

Summary of prior review in 2012

- The evidence supporting this measure consists of 11 nonrandomized studies that demonstrate that effective preventive measures for decreasing blood infection 'range from simple hand-washing protocols or closed medication delivery systems to more elaborate multidisciplinary quality improvement plans involving hand-washing, nutrition, skin care, respiratory care, vascular access, and diagnostic practices'.
- The 2012 Committee found this to be an important patient safety outcome measure.

Question for the Committee:

- *Is there at least one thing that the provider can do to achieve a change in the measure results?*

Guidance from algorithm #1 Evaluating Clinical Evidence: Health outcome or PRO (Box 1) → Relationship between the measure health outcome/ PRO and at least one health are action (Box 2) → Rate as pass.

Preliminary rating for evidence: ☒ **Pass** ☐ **No Pass**

1b. Gap in Care/Opportunity for Improvement and 1b. Disparities
Maintenance measures – increased emphasis on gap and variation

1b. Performance Gap. The performance gap requirements include demonstrating quality problems and opportunity for improvement.

The developer provides data from 2011 through 2013: in 2013 1,277 hospitals with a denominator population of 68,647 and 1,277 neonatal blood stream infections.

- The developer provides data files in the supplemental document including:

Distribution of Hospital Performance for NQI 03 Neonatal Blood Stream Infection Rate

Year	Number of Hospitals	Outcome of Interest	(Numerator)1 (Denominator) 1	Population at Risk Observed Rate Per 1000
2011	1,285	1,746	72,697	24.018
2012	1,344	1,695	74,032	22.896
2013	1,277	1,331	68,647	19.389

Disparities

The developer provides data from the Agency for Healthcare Research and Quality (AHRQ), Center for Delivery, Organization, and Markets, Healthcare Cost and Utilization Project, Nationwide Inpatient Sample, 2013, and AHRQ Quality Indicators.

Patient/hospital characteristic	Estimate	Std Error p-value (Ref Grp = *)	Lower 95% CL	Upper 95% CL
Gender				
Male*	19.4373	0.702		18.060
Female	17.132	0.727	0.011	15.708
Patient Zip Code Median Income				
First quartile (lowest income)	18.514	1.782	0.378	15.021
Second quartile	18.711	1.198	0.285	16.363
Third quartile	18.839	1.018	0.230	16.844
Fourth quartile (highest income)*	17.917	0.719		16.509
Location of patient residence (NCHS)				
Rural	17.774	5.044	0.456	7.888
Urban*	18.339	0.508		17.343
Expected payment source				
Private insurance*	16.681	0.813		15.087
Medicare1	7.898	12.131	0.235	0
Medicaid	19.575	0.681	0.003	18.240
Uninsured / self-pay / no charge	11.723	3.819	0.102	4.237
Other insurance	20.231	2.365	0.078	15.595
Location of Care				
Northeast*	17.577	1.335		14.963
Midwest	18.138	1.060	0.371	16.062
South	18.143	0.787	0.357	16.602
West	19.405	1.089	0.144	17.270

Source: Agency for Healthcare Research and Quality (AHRQ), Center for Delivery, Organization, and Markets, Healthcare Cost and Utilization Project, Nationwide Inpatient Sample, 2013, and AHRQ Quality Indicators, version 6.1 alpha.

Rates are adjusted by gender using the AHRQ QI PDI POA Reference Population for 2013 as the standard population

NCHS - National Center for Health Statistics designation for urban-rural locations.

1These births represent approximately 13,000 births covered by disabled Medicare beneficiaries.

Questions for the Committee:

- *Is there a gap in care that warrants a national performance measure?*
- *If no disparities information is provided, are you aware of evidence that disparities exist in this area of healthcare?*

Preliminary rating for opportunity for improvement: ☐ High ☒ Moderate ☐ Low ☐ Insufficient

Committee pre-evaluation comments

Criteria 1: Importance to Measure and Report (including 1a, 1b, 1c)

1a. Evidence *to Support Measure Focus*

Comments:

Yes. The health outcome is neonatal blood stream infection rate and the actions/interventions are hand hygiene and central line care.

Pass.

outcome measure

1b. Performance Gap

Comments:

Yes, performance data was provided from 2011-2013 from >1200 hospitals. Significant performance gap noted amongst various groups. Yes, subgroups were provided. Significant disparities noted amongst various subgroups (rural vs. urban, Medicaid vs. private insurance vs. uninsured, location of care). The data presented warrants this measure to be included as a national performance measure.

Moderate. Demonstrated disparities. Expands from the other measure which is tightly controlled to include a larger number of NICU admissions.

Data provided; high opportunity or improvement

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability

2a1. Reliability Specifications

Maintenance measures – no change in emphasis – specifications should be evaluated the same as with new measures

2a1. Specifications requires the measure, as specified, to produce consistent (reliable) and credible (valid) results about the quality of care when implemented.

Data source: Administrative claims Data

Specifications:

- The level of analysis is at the facility level and the care setting is hospital/acute care facility.
- The numerator includes; diagnosis of septicemia, diagnosis of newborn septicemia or bacteremia, and diagnosis of staphylococcal or Gram-negative bacterial infection.
- The denominator includes all newborns and outborns: with birth weight of 500 to 1,499 grams, gestational age between 24 and 30 weeks, birth weight greater than or equal to 1,500 grams who have died, birth weight greater than or equal to 1,500 grams with procedure codes for operating room procedure and or transferring from another health care facility within two days of birth.
- The denominator exclusions include: principle or secondary diagnosis of sepsis, sepsis or bacteria, staphylococcal or Gram-negative bacterial infection, birth weight less than 500 grams, length of stay less than 3 days, or with missing gender, age, quarter, year or principal diagnosis.
- An attached spreadsheet contains numerous ICD-9 and ICD-10 codes for [Numerator]: various septicemia diagnosis, newborn septicemia or bacteremia diagnosis, staphylococcal o gram-negative bacterial infection diagnosis, [Denominator]: gestational age, mechanical ventilation procedures, [Denominator Exclusions]: sepsis or bacteremia diagnosis, and for discharges in FY2012 and later.
- A calculation algorithm describes the process of calculating the measure.

Questions for the Committee:

- *Are all the data elements clearly defined? Are all appropriate codes included?*
- *Is the logic or calculation algorithm clear?*
- *Is it likely this measure can be consistently implemented?*

2a2. Reliability Testing Testing attachment
Maintenance measures – less emphasis if no new testing data provided

2a2. Reliability testing demonstrates if the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise enough to distinguish differences in performance across providers.

For maintenance measures, summarize the reliability testing from the prior review:

- Signal to noise analysis was performed using HCUP State Inpatient Databases (SID), Healthcare Cost and Utilization Project (HCUP) data from 2008, which includes 6 million pediatric discharges for 2,500 hospitals.
- The data demonstrated a systematic variation in the provider level rate of 0.419 to 69.167 per 1,000 from the 5th to 95th percentile and a signal ratio of 0.831 (outdated data).

Describe any updates to testing

- The developer added data from HCUP State Inpatient Databases (SID). Healthcare Cost and Utilization Project (HCUP). 2013. Agency for Healthcare Research and Quality, Rockville, MD. www.hcup-us.ahrq.gov/sidoverview.jsp.

SUMMARY OF TESTING

Reliability testing level ☒ **Measure score** ☐ **Data element** ☐ **Both**

Reliability testing performed with the data source and level of analysis indicated for this measure ☐ **Yes** ☒ **No**

Method of reliability testing

- Signal to noise was performed and in 2013 for 943 hospitals comprising on average 72.3 discharges per hospital, with a signal-to-noise average of .63456. Signal to noise is an appropriate method of assessing differences in measure performance.

Results of reliability testing

- Reliability increases with the number of observations:

Table 2. Signal-to-Noise Ratio by Size Decile for NQI 03 Neonatal Blood Stream Infection Rate

Size Decile	Number of Hospitals	Avg. Number of Discharges per Hospital in Decile	Avg. Signal-to-Noise Ratio for Hospitals in Decile
1	94	3.5	0.03776
2	94	6.5	0.04589
3	95	12.4	0.12740
4	94	20.0	0.19443
5	94	31.6	0.31406
6	95	48.0	0.40531
7	94	68.4	0.47076
8	95	101.3	0.58954
9	94	150.3	0.69027
10	94	281.9	0.79322
Overall	943	72.3	0.63456

Source: HCUP State Inpatient Databases (SID). Healthcare Cost and Utilization Project (HCUP). 2013. Agency for Healthcare Research and Quality, Rockville, MD. www.hcup-us.ahrq.gov/sidoverview.jsp. (AHRQ QI Software Version 6.1 alpha)

- The developer highlights that signal to noise ratios for hospitals with less than 32 qualifying discharges per year

was smaller. The developer notes that smoothed rates, which are recommended for all hospitals (and are implemented in the AHRQ software), address reliability concerns particularly for small hospitals. Hospitals with more than 48 discharges on average have risk adjusted rates with moderate to high reliability (average signal-to-noise ratio of 0.41 to 0.79). Overall, the signal to noise ratio for this indicator is good with an overall signal-to-noise ratio of 0.63.

Questions for the Committee:

- *Is the test sample adequate to generalize for widespread implementation?*
- *Do the results demonstrate sufficient reliability so that differences in performance can be identified?*

Preliminary rating for reliability: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

2b. Validity
Maintenance measures – less emphasis if no new testing data provided

2b1. Validity: Specifications

2b1. Validity Specifications. This section should determine if the measure specifications are consistent with the evidence.

Specifications consistent with evidence in 1a. ☒ **Yes** ☐ **Somewhat** ☐ **No**

Question for the Committee:

- *Are the specifications consistent with the evidence?*

2b2. Validity testing

2b2. Validity Testing should demonstrate the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality.

For maintenance measures, summarize the validity testing from the prior review:

Assessment of adequate face validity. The prior Committee did not voice any concerns.

Describe any updates to validity testing – none.

SUMMARY OF TESTING

Validity testing level ☒ **Measure score** ☐ **Data element testing against a gold standard** ☐ **Both**

Method of validity testing of the measure score:

- ☒ **Face validity only**
- ☐ **Empirical validity testing of the measure score**

Validity testing method:

Face validity was assessed using a multi-specialty panel. The rating scale rages from 1 – 9.

Question	Median	Agreement status
Overall rating of usefulness for internal QI improvement	8	Agree
Overall rating of usefulness for comparative purposes	8	Agree
Importance	8.5	Agree
Preventability	7	Indeterminate
Likelihood of medical error	6.5	Indeterminate
Charting	8	Agree

Bias	5	Indeterminate
Final recommendation		
Internal Quality Improvement: Recommend		
Comparative purposes: Recommend		

Validity testing results:

- This measure was deemed valid and the developer notes has high face validity for use in quality improvement and hospital comparative purposes.

Questions for the Committee:

- *Is the test sample adequate to generalize for widespread implementation?*
- *Do the results demonstrate sufficient validity so that conclusions about quality can be made?*
- *Do you agree that the score from this measure as specified is an indicator of quality?*

2b3-2b7. Threats to Validity

2b3. Exclusions:

The denominator exclusions include infants who do not meet eligibility following criteria:

- principal or secondary diagnosis present on admission for sepsis
- principal or secondary diagnosis present on admission for sepsis or bacteremia
- principal or secondary diagnosis present on admission for staphylococcal or Gram-negative bacterial infection
  - with birth weight less than 500 grams
  - with length of stay less than 3 days
  - with missing gender, age, quarter, year or principal diagnosis

- The developer does provide an analysis on the frequency of exclusions. Please see the chart below:

Table 3. Number and Percent of Discharges Excluded, by Denominator Exclusion Criteria for NQI 03 Neonatal Blood Stream Infection Rate[1]

NQI 03	Denominator			Potential Numerator[2]		
`Exclusion Name	Exclusion Count	After Exclusions	% Change	Exclusion Count	After Exclusions	% Change
Exclude Principal Diagnosis of Sepsis	350	2,988,719	0.01%	3	2,832	0.11%
Exclude birthweight < 500g	725	2,988,344	0.02%	50	2,785	1.76%
Exclude LOS < 3 days	19,967	2,969,102	0.67%	43	2,792	1.52%
No Exclusions applied	-	**2,989,069**	-	-	**2,835**	-

[1]This indicator does not have numerator exclusion criteria.

[2]Potential numerator cases are those that would have qualified for the numerator if not for a particular denominator exclusion criterion.

Source: HCUP State Inpatient Databases (SID). Healthcare Cost and Utilization Project (HCUP). 2013. Agency for Healthcare Research and Quality, Rockville, MD. www.hcup-us.ahrq.gov/sidoverview.jsp. (AHRQ QI Software Version 6.1 alpha)

The developer provides explanations for each exclusion:

- Patients with a principal diagnosis of sepsis are excluded because these infections are present on admission.
- Patients with a LOS of less than 3 days are excluded because sepsis cases in babies with a shorter LOS are likely to be perinatally acquired, and the transfer out rate is highest in the first 2 days of life. Patients with this short of a LOS have low risk of exposure to postnatally acquired sepsis.
- Patients with a birthweight of less than 500g are excluded from all NQI because these babies are very fragile and high risk; sepsis and other poor outcomes are less preventable in this population. Overall the exclusions account for less than 1% of denominator cases and less than 2% of numerator cases. Exclusions are retained to ensure face validity of the measure.

Questions for the Committee:

- *Are the exclusions consistent with the evidence?*
- *Are any patients or patient groups inappropriately excluded from the measure?*
- *Are the exclusions/exceptions of sufficient frequency and variation across providers to be needed (and outweigh the data collection burden)?*

2b4. Risk adjustment: **Risk-adjustment method** ☐ **None** ☒ **Statistical model** ☐ **Stratification**

Conceptual rationale for SDS factors included? ☐ **Yes** ☒ **No**

- Citing 'Late-onset sepsis in very low birth weight neonates: the experience of the NICHD Neonatal Research Network.' Stoll et al (2002) the there is no evidence or causal model to suggest that SDS factors are associated with late onset sepsis independent of quality of care, or are mediated by pre-hospital care.

SDS factors included in risk model? ☐ **Yes** ☒ **No**

Risk adjustment summary

- a standard set of covariates grouped into four categories: demographics, severity of illness, comorbidities and transfer-in status.
  Covariates that are considered as potential risk adjusters for the Neonatal Quality Indicators include age in days at admission, gender, birthweight, Major Diagnostic Categories (MDC), MS-DRG (removing comorbidity and complication distinctions), availability of data elements (point of origin and procedure dates) and whether they were transferred from another facility.
- Once the preliminary multivariable model is specified, it is estimated on the pediatric analytic data, as appropriate using a backward stepwise bootstrap approach.

Additional details are available in the AHRQ Quality Indicator Empirical Methods document, included in the supplemental materials.

Table 4. Risk Adjustment Coefficients for NQI 03 Neonatal Blood Stream Infection Rate

Parameter	Label	DF	Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square
Intercept	Intercept	1	-3.2646	0.3193	104.5576	<.0001
PD_MALE_1c	Male | Age Days = 0	1	0.1484	0.0544	7.4351	0.0064
PD_MALE_2c	Male | Age Days >= 1	1	0.0581	0.1694	0.1175	0.7317
PD_FEMALE_	Female | Age Days >= 1	1	0.0594	0.1963	0.0917	0.762
BWHTCAT_1	Birth Weight 2500+ grams	1	-0.2312	0.1769	1.7088	0.1911
BWHTCAT_2	Birth Weight 2000 to 2499 grams	1	-0.0139	0.1546	0.0081	0.9283
BWHTCAT_3	Birth Weight 1750 to 1999 grams	1	0.3669	0.1722	4.5401	0.0331

BWHTCAT_4	Birth Weight 1500 to 1749 grams	1	0.3944	0.1525	6.6837	0.0097
BWHTCAT_5	Birth Weight 1250 to 1499 grams	1	0.3083	0.1241	6.1738	0.013
BWHTCAT_6	Birth Weight 1000 to 1249 grams	1	1.2207	0.1078	128.1888	<.0001
BWHTCAT_7	Birth Weight 750 to 999 grams	1	1.9086	0.1003	362.023	<.0001
BWHTCAT_8	Birth Weight <= 749 grams	1	2.3741	0.1003	560.461	<.0001
BWHTCAT_1b	Birth Weight <= 749 grams	0	0	0	.	<.0001
BWHTCAT_2b	Birth Weight > 749 grams	0	0	0	.	<.0001

Source: HCUP State Inpatient Databases (SID). Healthcare Cost and Utilization Project (HCUP). 2013. Agency for Healthcare Research and Quality, Rockville, MD. www.hcup-us.ahrq.gov/sidoverview.jsp. (AHRQ QI Software Version 6.1 alpha)

Parameter	Label	DF	Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square
MDC_1	MDC 1: Nervous System	1	-2.6364	1.0481	6.3267	0.0119
MDC_11	MDC 11: Kidney And Urinary Tract	1	-1.3073	1.0534	1.5402	0.2146
MDC_15	MDC 15: Newborn And Other Neonates	1	-1.669	0.3151	28.0491	<.0001
MDC_3	MDC 3: Ear Nose Mouth And Throat	1	-0.8865	0.7867	1.2698	0.2598
MDC_4	MDC 4: Respiratory System	1	-0.9382	0.4602	4.1572	0.0415
MDC_5	MDC 5: Circulatory System	1	-0.8543	0.3104	7.573	0.0059
MDC_6	MDC 6: Digestive System	1	-1.015	0.3717	7.455	0.0063
MDC_8	MDC 8: Musculoskeletal And Connective	1	-0.8028	0.5857	1.8789	0.1705
NOPOUB04	No Point Of Origin	1	-0.1471	0.0803	3.3556	0.067
NOPRDAY	Days to procedure flag	1	-0.8909	0.1928	21.3446	<.0001
mdrg_103	Craniotomy and endovascular intracranial procedure	1	1.4721	1.4209	1.0733	0.3002
mdrg_7799	ECMO or trach w MV 96+ hrs or PDX exc face	1	1.427	0.1394	104.7977	<.0001
mdrg_8898	Ungroupable	1	0.5186	0.2665	3.7882	0.0516
c-statistic=0.752						

Source: HCUP State Inpatient Databases (SID). Healthcare Cost and Utilization Project (HCUP). 2013. Agency for Healthcare Research and Quality, Rockville, MD. www.hcup-us.ahrq.gov/sidoverview.jsp. (AHRQ QI Software Version 6.1 alpha)

Questions for the Committee:

- *Is an appropriate risk-adjustment strategy included in the measure?*
- *Are the candidate and final variables included in the risk adjustment model adequately described for the measure to*

be implemented?

- *Are all of the risk adjustment variables present at the start of care? If not, describe the rationale provided.*
- *Do you agree with the developer's rationale that there is no conceptual basis for adjusting this measure for SDS factors?*
- *Do you agree with the developer's decision, based on their analysis, to not include SDS factors in their risk-adjustment model?*

2b5. Meaningful difference *(can statistically significant and clinically/practically meaningful differences in performance measure scores can be identified)*:

- The developer provides a table of Performance Categories by Hospital Size Decile for NQI 03 Neonatal Blood Stream Infection Rate.
- The developer concludes "This indicator has poor discrimination for most hospitals to identify low performing hospitals; 11% of hospitals can be classified as better or worse than the threshold (the percentage classified as either above or below the threshold). The indicator has poor discrimination, particularly for moderate to large hospitals to identify high performing hospitals; 22% of hospitals can be classified as better or worse than the benchmark. These results reflect the skewed distribution for this indicator, due to many hospitals with very small denominators and zero numerators. For the largest hospitals with 159 or more high risk infants in the denominator, performance discrimination improves with about 55% of hospitals classified as better or worse than the benchmark and 35% - 47% hospitals classified as better or worse than the threshold."

Question for the Committee:

- *Does this measure identify meaningful differences about quality?*

2b6. Comparability of data sources/methods:

N/A

2b7. Missing Data

- The AHRQ QIs use frequently reported administrative data variables. NQI 03 excludes cases with missing age, sex, discharge quarter, discharge year, and principal diagnosis. These variables are required for indicator construction and are required of all hospital discharge records. The rate of missing data for each variable is available by state and year from the AHRQ HCUP website (http://www.hcup-us.ahrq.gov/cdstats/cdstats_search.jsp). For these variables, rates of missing data are typically less than 1% of the state database. It is unlikely the bias would occur from such a low rate of missing data.

Guidance from the algorithm:
Consistent with specifications (Box 1) →potential threats to validity addressed (Box 2) →face validity systematically assessed (Box 4) → substantial agreement (Box 5) → moderate (in the absence of empirical validity testing, moderate is the highest rating possible)

Preliminary rating for validity: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 2: Scientific Acceptability of Measure Properties (including all 2a, 2b, and 2d)

2a1. & 2b1. Specifications

Comments:

**No major concerns about reliability-specifications, although small hospitals with <32 qualifying discharges per year

had a higher signal-to-noise ratio.**
Data elements clear. Supports moderate.
All defined in ICD 10 per developer. Risk-adjusted measure appears clearly defined.
No major concerns about validity.
Supports moderate.
In the data provided, it appears consistent.

2a2. Reliability Testing
Comments:
Yes, reliability testing was performed using signal-to-noise analysis of HCUP data from 2008. Additional updated data available from 2013.
**Score. Supports moderate. **
Moderate reliability (in the absence of testing the measure score.)

2b2. Validity Testing
Comments:
Face validity testing done. I am not familiar with validity-testing methodologies.
High face validity.
Correlation tables and scatter plots were reviewed; appear to reflect an indicator of quality.

2b3. Exclusions Analysis
2b4. Risk Adjustment/Stratification for Outcome or Resource Use Measures
2b5. Identification of Statistically Significant & Meaningful Differences In Performance
2b6. Comparability of Performance Scores When More Than One Set of Specifications
2b7. Missing Data Analysis and Minimizing Bias
Comments:
**Extremely premature infants with birth weight <500g have been excluded from the analysis. I agree that the number of these infants is extremely low and acceptable to exclude them currently. No patients seem to be inappropriately excluded. SDS factor was not included in risk model. **
Statistical model, no SDS factors (2002, + disparities? in related morbidities), poor discrimination for most hospitals to identify low performance/moderate, + threats (moderate/low).
If no data, assigned an E

Criterion 3. Feasibility
Maintenance measures – no change in emphasis – implementation issues may be more prominent

3. Feasibility is the extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

- This measure is based on administrative claims. Claims data is generally thought to be of minimal burden and highly feasible.

Questions for the Committee:
- *Are the required data elements routinely generated and used during care delivery?*
- *Are the required data elements available in electronic form, e.g., EHR or other electronic sources?*
- *Is the data collection strategy ready to be put into operational use?*

Preliminary rating for feasibility: ☒ **High** ☐ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments
Criteria 3: Feasibility

3a. Byproduct of Care Processes
3b. Electronic Sources
3c. Data Collection Strategy
Comments:
No concerns about operational use of data from available sources. These data are routinely generated and commonly used during infant's care.
High/moderate
Moderate- currently reported to Joint commission.

Criterion 4: Usability and Use
Maintenance measures – increased emphasis – much greater focus on measure use and usefulness, including both impact /improvement and unintended consequences

4. Usability and Use evaluate the extent to which audiences (e.g., consumers, purchasers, providers, policymakers) use or could use performance results for both accountability and performance improvement activities.

Current uses of the measure
Publicly reported? ☒ **Yes** ☐ **No**

Current use in an accountability program? ☒ **Yes** ☐ **No**
OR
Planned use in an accountability program? ☐ **Yes** ☐ **No**

Accountability program details

- Wisconsin Hospital Association (WHA) Information Center
- Wisconsin Hospital Association (WHA) Quality Indicators Report
  http://www.whainfocenter.com/uploads/PDFs/Publications/QualityIndicators/2012_WI_IQIReport.pdf

Improvement results

- The developer states that "Overall and average hospital performance has improved from 24.0 in 2011 to 19.4 in 2013 nationwide. For hospitals, the average rate has also decreased from 11.5 in 2011 to 9.15 in 2013, and the variation in hospital rates has also decreased with a standard deviation of 22.6 vs. 19.1."

Unexpected findings (positive or negative) during implementation
None reported

Potential harms
None reported

Feedback:

***Questions for the Committee**:*
- *Committee members should share any experience with use of this measure.*
- *What does experience with this measure tell various stakeholders?*
- *How can the performance results be used to further the goal of high-quality, efficient healthcare?*
- *Do the benefits of the measure outweigh any potential unintended consequences?*

Preliminary rating for usability and use: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 4: Usability and Use

4a. Accountability and Transparency

4b. Improvement

4c. Unintended Consequences

Comments:

**This measure is being currently used for public reporting and in accountability programs. There are currently 2 other NQF measures that are closely related to this measure (0304 and 1731). Harmonization with these measures would be beneficial. **

**Supports moderate. **

Already in use. publicly reported.

Criterion 5: Related and Competing Measures

Related or competing measures

- 1731: PC-04 Health Care-Associated Bloodstream Infections in Newborns

Harmonization

- The developer states the two measures are harmonized to the extent possible given different data sources.
- IN the 2012 evaluation the Committee noted that different data sources are important for different users: states, Medicaid and purchasers do not have access to chart data and rely on administrative data. Chart based measures provide more clinical detail for feedback/quality improvement.

Pre-meeting public and member comments

-

NATIONAL QUALITY FORUM

Measure missing data in MSF 6.5 from MSF 5.0

NQF #: 0478	**NQF Project:** Perinatal and Reproductive Health Project

1. IMPACT, OPPORTUITY, EVIDENCE - IMPORTANCE TO MEASURE AND REPORT

Importance to Measure and Report is a threshold criterion that must be met in order to recommend a measure for endorsement. All three subcriteria must be met to pass this criterion. See guidance on evidence.

Measures must be judged to be important to measure and report in order to be evaluated against the remaining criteria. **(evaluation criteria)**

1c.1 Structure-Process-Outcome Relationship *(Briefly state the measure focus, e.g., health outcome, intermediate clinical outcome, process, structure; then identify the appropriate links, e.g., structure-process-health outcome; process-health outcome; intermediate clinical outcome-health outcome)***:**

Effective preventive measures range from simple hand-washing protocols or closed medication delivery systems to more elaborate multidisciplinary quality improvement plans involving hand-washing, nutrition, skin care, respiratory

care, vascular access, and diagnostic practices. All of these interventions have been shown to substantially reduce infection rates, albeit in nonrandomized studies using historical or concurrent control units (Adams-Chapman & Stoll, 2002; Aly et al., 2005; Bloom et al., 2003; Clark et al., 2004a; Clark et al., 2004b; Horbar et al., 2001; Lam et al., 2004; Kilbride et al., 2003a; Kilbride et al., 2003b; Ng et al., 2004; Schelonka et al., 2006). For example, six Vermont Oxford Network NICUs reduced their rates of coagulase-negative staphylococcus infections from 22.0% in 1994 to 16.6% in 1996 after implementing a quality improvement model (versus a much smaller decrease from 15.4% to 14.5% at 66 comparison NICUs) (Horbar et al., 2001). A similar reduction from 24.6% to 16.4% was achieved with a multi-modality, multi-hospital intervention focusing on hand hygiene with an effective agent before and after every patient contact, eliminating hand jewelry and artificial nails, using maximal barrier precautions during central venous catheter insertion, decreasing the number of skin punctures, reducing the duration of intravenous lipid and deep line use, and improving the diagnosis of health care-associated infections. (Kilbride et al., 2003a; Kilbride et al., 2003b).

Given the fragility and susceptibility of the patient population, a baseline level of health care-associated infections will be expected, even with good protocols in place. However, those centers that have prevention protocols, and are able to encourage health care workers to adhere to these protocols, will probably have success in reducing their rates of health care-associated bacteremia in their neonatal population. Indeed, several quasi-experimental studies have demonstrated that NICUs can lower their infection rates (based on positive blood cultures) from as high as 13.5 per 1,000 patient days to as low as 3.0 per 1,000 patient days(Adams-Chapman & Stoll, 2002; Aly et al., 2005; Bloom et al. ,2003; Clark et al., 2004a; Clark et al., 2004b; Horbar et al., 2001; Lam et al., 2004; Kilbride et al., 2003a; Kilbride et al., 2003b; Ng et al., 2004; Schelonka et al., 2006).

1c.2-3 Type of Evidence *(Check all that apply)***:**

Selected individual studies (rather than entire body of evidence)

1c.4 **Directness of Evidence to the Specified Measure** *(State the central topic, population, and outcomes addressed in the body of evidence and identify any differences from the measure focus and measure target population)***:**

Not applicable

1c.5 **Quantity of Studies in the Body of Evidence** *(Total number of studies, not articles*)**:** Not applicable

1c.6 **Quality of Body of Evidence** (*Summarize the certainty or confidence in the estimates of benefits and harms to patients across studies in the body of evidence resulting from study factors. Please address: a) study design/flaws; b) directness/indirectness of the evidence to this measure (e.g., interventions, comparisons, outcomes assessed, population included in the evidence); and c) imprecision/wide confidence intervals due to few patients or events):* Not applicable

1c.7 **Consistency of Results across Studies** *(Summarize the consistency of the magnitude and direction of the effect):* Not applicable

1c.8 **Net Benefit** *(Provide estimates of effect for benefit/outcome; identify harms addressed and estimates of effect; and net benefit - benefit over harms)***:**

Not applicable

1c.9 **Grading of Strength/Quality of the Body of Evidence**. Has the body of evidence been graded? No

1c.10 **If body of evidence graded, identify the entity that graded the evidence including balance of representation and any disclosures regarding bias:** Not applicable

1c.11 **System Used for Grading the Body of Evidence:** Other

1c.12 **If other, identify and describe the grading scale with definitions:** Not applicable

1c.13 **Grade Assigned to the Body of Evidence:** Not applicable

1c.14 **Summary of Controversy/Contradictory Evidence:** Not applicable

1c.15 **Citations for Evidence other than Guidelines*(Guidelines addressed below)*:**

Adams-Chapman, I. & Stoll, B.J. (2002). Prevention of nosocomial infections in the neonatal intensive care unit. Current Opinion in Pediatrics.14 (2):157-64.

Aly, H., Herson, V., Duncan, A., et al. (2005). Is bloodstream infection preventable among premature infants? A tale of two

cities. Pediatrics. 115(6):1513-8.

Bloom, B.T., Craddock, A., Delmore, P.M., et al. (2003). Reducing acquired infections in the NICU: observing and implementing meaningful differences in process between high and low acquired infection rate centers. Journal of Perinatology. 23(6):489-92.

Clark, R., Powers, R., White, R., Bloom, B., Sanchez, P., & Benjamin, D.K., Jr. (2004a). Prevention and treatment of nosocomial sepsis in the NICU. Journal of Perinatology. 4; 24(7):446-53.

Clark, R., Powers, R., White, R., Bloom, B., Sanchez, P., & Benjamin, D.K., Jr. (2004b). Nosocomial infection in the NICU: a medical complication or unavoidable problem? Journal of Perinatology. 24(6):382-8.

Horbar, J.D., Rogowski, J., Plsek, P.E., et al. (2001). Collaborative quality improvement for neonatal intensive care. NIC/Q Project Investigators of the Vermont Oxford Network. Pediatrics. 107(1):14-22.

Kilbride, H.W., Wirtschafter, D.D., Powers, R.J., & Sheehan, M.B. (2003a). Implementation of evidence-based

potentially better practices to decrease nosocomial infections. Pediatrics. 111(4 Pt 2):e519-33.

Kilbride, H.W., Powers, R., Wirtschafter, D.D., et al. (2003b). Evaluation and development of potentially better practices to prevent neonatal nosocomial bacteremia. Pediatrics. 111(4 Pt 2):e504-18.

Lam, B.C., Lee, J., & Lau, Y.L. (2004). Hand Hygiene Practices in a Neonatal Intensive Care Unit: A Multimodal Intervention and Impact on Nosocomial Infection. Pediatrics.114 (5):e565.

Ng, P.C., Wong, H.L., Lyon, D.J., et al. (2004). Combined use of alcohol hand rub and gloves reduces the incidence of late onset infection in very low birthweight infants. Archives of Disease in Childhood Fetal &

Neonatal Edition. 89(4):F336-40.

Schelonka, R.L., Scruggs, S., Nichols, K., Dimmitt, R.A., & Carlo, W.A. (2006). Sustained reductions in neonatal nosocomial infection rates following a comprehensive infection control intervention. Journal of Perinatology. 26(3):176-9.

1c.16 Quote verbatim, <u>the specific guideline recommendation</u> *(Including guideline # and/or page #)***:**

Not applicable

1c.17 Clinical Practice Guideline Citation: Not applicable

1c.18 National Guideline Clearinghouse or other URL: Not applicable

1c.19 Grading of Strength of Guideline Recommendation. Has the recommendation been graded? No

1c.20 If guideline recommendation graded, identify the entity that graded the evidence including balance of representation and any disclosures regarding bias:

1c.21 System Used for Grading the Strength of Guideline Recommendation: Other

1c.22 If other, identify and describe the grading scale with definitions: Not applicable

1c.23 Grade Assigned to the Recommendation: Not applicable

1c.24 Rationale for Using this Guideline Over Others: Not applicable

Based on the NQF descriptions for rating the evidence, what was the <u>developer's assessment</u> of the quantity, quality, and consistency of the body of evidence?

1c.25 Quantity: Moderate **1c.26** Quality: Moderate**1c.27** Consistency**:** Moderate

1. Evidence, Performance Gap, Priority – Importance to Measure and Report

Extent to which the specific measure focus is evidence-based, important to making significant gains in healthcare quality, and improving health outcomes for a specific high-priority (high-impact) aspect of healthcare where there is variation in or overall less-than-optimal performance. ***Measures must be judged to meet all subcriteria to pass this criterion and be evaluated against the remaining criteria.***

1a. Evidence to Support the Measure Focus – See attached Evidence Submission Form

0478_Evidence_MSF5.0_Data-635787040715945660.doc

1b. Performance Gap

Demonstration of quality problems and opportunity for improvement, i.e., data demonstrating:

- considerable variation, or overall less-than-optimal performance, in the quality of care across providers; and/or
- disparities in care across population groups.

1b.1. **Briefly explain the rationale for this measure** *(e.g., the benefits or improvements in quality envisioned by use of this measure)*

Low birth weight, or premature, and critically ill infants are at increased risk for sepsis or blood stream infections due to immature immune systems, immature skin barriers, and invasive devices such as central venous access or arterial access, ventilation or feeding tubes. Septicemia is one of the most common neonatal infection in neonatal intensive care units. Processes such as hand washing, evidence based vascular access procedures and central line care, and appropriate administration of prophylactic antibiotics can lower rates of neonatal blood stream infection.

1b.2. **Provide performance scores on the measure as specified (current and over time) at the specified level of analysis.** *(This is required for endorsement maintenance. Include mean, std dev, min, max, interquartile range, scores by decile. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included). This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*

This table is also included in the supplemental files.

Table 1. Reference Population Rate and Distribution of Hospital Performance for NQI 03 Neonatal Blood Stream Infection Rate

Overall Reference Population Rate

Year3	Number of Hospitals (Numerator)1	Outcome of Interest Population at Risk (Denominator)1	Observed Rate Per 10001
2011	1,285	1,746	72,697	24.018
2012	1,344	1,695	74,032	22.896
2013	1,277	1,331	68,647	19.389

Distribution of Hospital-level Observed Rates in Reference Population Per 1000

Year3	Number of Hospitals	(p=percentile)2 Mean	SD	p5	p25	Median	p75	p95
2011	1,285	11.53	22.62	0.00	0.00	0.00	16.81	56.14
2012	1,344	11.62	26.61	0.00	0.00	0.00	14.04	58.82
2013	1,277	9.15	19.05	0.00	0.00	0.00	11.83	49.28

Source: HCUP State Inpatient Databases (SID). Healthcare Cost and Utilization Project (HCUP). 2011-2013. Agency for Healthcare Research and Quality, Rockville, MD. www.hcup-us.ahrq.gov/sidoverview.jsp. (AHRQ QI Software Version 6.1 alpha)

1The observed rate refers to the total rate for all observations included in the reference population data (numerator) divided by the total combined eligible population of all hospitals included in the reference population data (denominator).

2The distribution of hospital rates reports the mean and standard deviation (SD) of the observed rates for all hospitals included in the dataset, as well as the observed rate for hospitals in the 5th, 25th, 50th (median), 75th, and 95th percentile.

3 Reference population is limited to states with present on admission data (POA). Since many states did not report POA data prior to 2011 we have not included testing prior to 2011.

1b.3. **If no or limited performance data on the measure as specified is reported in 1b2, then provide a summary of data from the literature that indicates opportunity for improvement or overall less than optimal performance on the specific focus of measurement.**

n/a

1b.4. **Provide disparities data from the measure as specified (current and over time) by population group, e.g., by race/ethnicity, gender, age, insurance status, socioeconomic status, and/or disability.** *(This is required for endorsement maintenance. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities include.) This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*
This table is also included in the supplemental files.
Table 2. Neonatal Blood Stream Infection Rate per 1,000 (NQI 03), by patient and hospital characteristics, 2013

Patient/hospital characteristic	Estimate	Std Error	p-value	(Ref Grp = *)	Lower 95% CL	Upper 95% CL
Total U.S.	18.3284	0.505		17.339	19.318	
Patient Characteristics						
Gender:						
Male*	19.4373	0.702		18.060	20.814	
Female	17.132	0.727	0.011	15.708	18.556	
Patient Zip Code Median Income						
First quartile (lowest income)			18.514	1.782	0.378	15.021	22.008
Second quartile	18.711	1.198	0.285	16.363	21.059	
Third quartile	18.839	1.018	0.230	16.844	20.835	
Fourth quartile (highest income)*			17.917	0.719		16.509	19.326
Location of patient residence (NCHS):						
Rural	17.774	5.044	0.456	7.888	27.659	
Urban*	18.339	0.508		17.343	19.335	
Expected payment source:						
Private insurance*			16.681	0.813		15.087	18.275
Medicare1	7.898	12.131	0.235	0	31.675	
Medicaid	19.575	0.681	0.003	18.240	20.912	
Uninsured / self-pay / no charge			11.723	3.819	0.102	4.237	19.208
Other insurance	20.231	2.365	0.078	15.595	24.867	
Location of Care:						
Northeast*			17.577	1.335		14.963	20.193
Midwest	18.138	1.060	0.371	16.062	20.215	
South	18.143	0.787	0.357	16.602	19.685	
West	19.405	1.089	0.144	17.270	21.540	

Source: Agency for Healthcare Research and Quality (AHRQ), Center for Delivery, Organization, and Markets, Healthcare Cost and Utilization Project, Nationwide Inpatient Sample, 2013, and AHRQ Quality Indicators, version 6.1 alpha.
Rates are adjusted by gender using the AHRQ QI PDI POA Reference Population for 2013 as the standard population
NCHS - National Center for Health Statistics designation for urban-rural locations.
1These births represent approximately 13,000 births covered by disabled Medicare beneficiaries.

1b.5. **If no or limited data on disparities from the measure as specified is reported in 1b4, then provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement. Include citations.**
n/a

1c. High Priority (previously referred to as High Impact)

The measure addresses:

- a specific national health goal/priority identified by DHHS or the National Priorities Partnership convened by NQF; OR
- a demonstrated high-priority (high-impact) aspect of healthcare (e.g., affects large numbers of patients and/or has a substantial impact for a smaller population; leading cause of morbidity/mortality; high resource use (current and/or future); severity of illness; and severity of patient/societal consequences of poor quality).

1c.1. **Demonstrated high priority aspect of healthcare**
Patient/societal consequences of poor quality
1c.2. **If Other:**

1c.3. **Provide epidemiologic or resource use data that demonstrates the measure addresses a high priority aspect of healthcare. List citations in 1c.4.**
In the 2013 AHRQ QI POA Reference Population (34 State Inpatient Databases (SID)) there were 1,331 qualifying blood stream

infection events out of 68,647 high risk newborns, for a rate of 19.4 per 1,000 (see Table 1, Section 1b.2).

Health care-associated bacteremia is a significant problem for infants admitted into neonatal intensive care units (NICUs) and other hospital units. This is especially true for very low birth weight infants who are at high risk for these infections due to their immature immune systems, immature skin barriers and need for invasive monitoring and supportive care (Adams-Chapman & Stoll, 2002; Bloom et al., 2003; Clark et al., 2004a; Clark et al., 2004b; Gaynes et al., 1996; Payne et al., 2004; Sohn et al., 2001; Stoll et al., 2002). Reported health care-associated infection rates range from 6% to 33%, but the rate varies widely among different centers (Adams-Chapman & Stoll, 2002; Bloom et al.,2003; Clark et al.,2004b; Sohn et al.,2001; Stoll et al.,2002). Neonatal blood stream infections have been associated with adverse outcomes such as increased mortality, cost and length of stay (Adams-Chapman & Stoll, 2002; Bloom et al.,2003; Clark et al., 2004b; Horbar et al., 2001; Kilbride et al., 2003a; Sohn et al.,2001; Stoll et al., 2002, Donavan et al. 2013). Donovan et al. (2013) found that very low birthweight infants in three Cincinnati area NICUs with hospital-associated blood stream infection had greater costs (without physician charges) and length of stay (LOS) as compared to infants without infection, adjusted for clinical risk factors present during the first 3 days of life ($106,621 vs. $158,672 and LOS 46 vs 56, p<.05). The increase in cost and LOS for infants with a blood stream infection was most pronounced among infants without a major anomaly or surgery, who survived to NICU discharge, but the increased risk remained for those that died prior to discharge. The mortality rate prior to NICU discharge for those with infection was 30% compared with 7% for infants with no infection (chi-square, $p < 0.0001$). Stoll et al. (2002) examined registry data from the NICHD Research Network. Infants with late-onset sepsis had longer length of stay than those without (79 v 60 days, $p < 0.001$), and higher mortality, controlling for gestational age and sex (18% v 7%, $p < 0.001$).

1c.4. Citations for data demonstrating high priority provided in 1a.3

HCUP State Inpatient Databases (SID). Healthcare Cost and Utilization Project (HCUP). 2009-2013. Agency for Healthcare Research and Quality, Rockville, MD. www.hcup-us.ahrq.gov/sidoverview.jsp. (AHRQ QI Software Version 6.1 alpha)
Adams-Chapman, I. & Stoll, B.J. (2002). Prevention of nosocomial infections in the neonatal intensive care unit. Current Opinion in Pediatrics.14 (2):157-64.
Barton, L., Hodgman, J.E., & Pavlova, Z. (1999). Causes of death in the extremely low birth weight infant. Pediatrics. 103(2):446-51.
Bloom, B.T., Craddock, A., Delmore, P.M., et al. (2003). Reducing acquired infections in the NICU: observing and implementing meaningful differences in process between high and low acquired infection rate centers. Journal of Perinatology. 23(6):489-92.
Clark, R., Powers, R., White, R., Bloom, B., Sanchez, P., & Benjamin, D.K., Jr. (2004a). Prevention and treatment of nosocomial sepsis in the NICU. Journal of Perinatology. 4; 24(7):446-53.
Clark, R., Powers, R., White, R., Bloom, B., Sanchez, P., & Benjamin, D.K., Jr. (2004b). Nosocomial infection in the NICU: a medical complication or unavoidable problem? Journal of Perinatology. 24(6):382-8.
Gaynes, R.P., Edwards, J.R., Jarvis, W.R., Culver, D.H., Tolson, J.S., & Martone, W.J. (1996). Nosocomial infections among neonates in high-risk nurseries in the United States. National Nosocomial Infections Surveillance System. Pediatrics. 98(3 Pt 1):357-61.
Horbar, J.D., Rogowski, J., Plsek, P.E., et al. (2001). Collaborative quality improvement for neonatal intensive care. NIC/Q Project Investigators of the Vermont Oxford Network. Pediatrics. 107(1):14-22.
Kilbride, H.W., Wirtschafter, D.D., Powers, R.J., & Sheehan, M.B. (2003a). Implementation of evidence-based potentially better practices to decrease nosocomial infections. Pediatrics. 111(4 Pt 2):e519-33.
Payne, N.R., Carpenter, J.H., Badger, G.J., Horbar, J.D., & Rogowski, J. (2004). Marginal increase in cost and excess length of stay associated with nosocomial bloodstream infections in surviving very low birth weight infants. Pediatrics. 114(2):348-55.
Sohn, A.H., Garrett, D.O., Sinkowitz-Cochran, R.L., et al. (2001). Prevalence of nosocomial infections in neonatal intensive care unit patients: Results from the first national point-prevalence survey. Journal of Pediatrics. 139(6):821-7.
Stoll, B.J., Hansen, N., Fanaroff, A.A., et al. (2002). Late-onset sepsis in very low birth weight neonates: the experience of the NICHD Neonatal Research Network. Pediatrics. 110(2 Pt 1):285-91.

1c.5. If a PRO-PM (e.g. HRQoL/functional status, symptom/burden, experience with care, health-related behaviors), provide evidence that the target population values the measured PRO and finds it meaningful. *(Describe how and from whom their input was obtained.)*

n/a

2. Reliability and Validity—Scientific Acceptability of Measure Properties

Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. ***Measures must be judged to meet the subcriteria for both reliability and validity to pass this criterion and be evaluated against the remaining criteria.***

2a.1. Specifications The measure is well defined and precisely specified so it can be implemented consistently within and across organizations and allows for comparability. eMeasures should be specified in the Health Quality Measures Format (HQMF) and the Quality Data Model (QDM).

De.5. **Subject/Topic Area** *(check all the areas that apply):*
Perinatal and Reproductive Health

De.6. **Cross Cutting Areas** *(check all the areas that apply):*
Safety : Healthcare Associated Infections

S.1. **Measure-specific Web Page** *(Provide a URL link to a web page specific for this measure that contains current detailed specifications including code lists, risk model details, and supplemental materials. Do not enter a URL linking to a home page or to general information.)*
1.usa.gov/1ZCFktQ Note: The URL link currently provides Version 5.0 specifications. Version 6.0 specifications will be released publicly March 2016 and found via the module page: http://www.qualityindicators.ahrq.gov/Modules/pdi_resources.aspx

S.2a. **If this is an eMeasure**, HQMF specifications must be attached. Attach the zipped output from the eMeasure authoring tool (MAT) - if the MAT was not used, contact staff. (Use the specification fields in this online form for the plain-language description of the specifications)
This is not an eMeasure **Attachment:**

S.2b. **Data Dictionary, Code Table, or Value Sets** *(and risk model codes and coefficients when applicable) must be attached. (Excel or csv file in the suggested format preferred - if not, contact staff)*
Attachment **Attachment:** NQI03_Tech_Specs_v6.1alpha_160211xlsx.xlsx

S.3. **For endorsement maintenance**, please briefly describe any changes to the measure specifications since last endorsement date and explain the reasons.
As standard protocol, the AHRQ QI program annually updates all measures with Fiscal Year coding changes, refinements based on stakeholder input, refinements to improve specificity and sensitivity based on additional analyses, and necessary software changes. In addition, approximately every two years, AHRQ updates the risk adjustment parameter estimates and composite weights based on the most recent year of data (i.e., the most current reference population possible). The refined measures are tested and confirmed to be valid and reliable prior to release of the updated software.

Since the last update, the following changes have been made to the indicator:

Changes to Numerator
Added code to numerator (septicemia):
Added: 3812 – METH RES STAPH AUR SEPT
Added: 382 – PNEUMOCOCCAL SEPTICEMIA [STREPTOCOCCUS PNEUMONIAE SEPTICEMIA]
Added: 388 – OTHER SPECIFIED SEPTICEMIAS

Removed code from numerator (newborn septicemia or bacteremia diagnosis codes):
Removed: 7907 BACTEREMIA
Denominator Exclusions
Changed length of stay exclusion from <2 days to <3 days. Note that a version 4.5 and 5.0 had a different LOS cutoff (LOS <9 days). Change was made for harmonization purposes.
Added codes to denominator exclusion, Sepsis or bacteremia diagnosis codes:
Added: 1125 – DISSEMINATED CANDIDIASIS
Added: 77181 – NB SEPTIEMIA [SEPSIS]
Added: 77183 – BACTEREMIA OF NEW BORN
Added: 7907 – BACTEREMIA
For discharges in FY2012 and later: (SEPTI2D)

Removed: 381 – STAPHYLOCOCCAL SEPTICEMIA
Added: 9980 – POSTOPERATIVE SHOCK / * NOT VALID AFTER OCTOBER 1, 2011
Additional information regarding revisions to PDI software and technical specifications visit:
http://www.qualityindicators.ahrq.gov/Modules/pdi_resources.aspx
Note: Version 6.0 specifications will be released publicly March 2016.

S.4. **Numerator Statement** *(Brief, narrative description of the measure focus or what is being measured about the target population, i.e., cases from the target population with the target process, condition, event, or outcome)*
IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
Discharges, among cases meeting the inclusion and exclusion rules for the denominator, with either:

- any secondary ICD-9-CM or ICD-10 CM diagnosis codes for other septicemia; or
- any secondary ICD-9-CM or ICD-10 CM diagnosis codes for newborn septicemia or bacteremia and
- any secondary ICD-9-CM or ICD-10 CM diagnosis codes for staphylococcal or Gram-negative bacterial infection

S.5. **Time Period for Data** *(What is the time period in which data will be aggregated for the measure, e.g., 12 mo, 3 years, look back to August for flu vaccination? Note if there are different time periods for the numerator and denominator.)*
The time period is one year for users with a complete sample of hospital discharges (i.e., "all payer" data). Note that the signal variance parameters assume a one-year time period. Users may use longer time periods if desired.

S.6. **Numerator Details** *(All information required to identify and calculate the cases from the target population with the target process, condition, event, or outcome such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
IF an OUTCOME MEASURE, describe how the observed outcome is identified/counted. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
Please see attached excel file in S.2b. for version 6.1 alpha specifications.

S.7. **Denominator Statement** *(Brief, narrative description of the target population being measured)*
All newborns and outborns with either:

- a birth weight of 500 to 1,499 grams (Birth Weight Categories 2, 3, 4 and 5); or
- any-listed ICD-9-CM or ICD-10 CM diagnosis codes for gestational age between 24 and 30 weeks; or
- a birth weight greater than or equal to 1,500 grams (Birth Weight Category 6, 7, 8, or 9) and death (DISP=20); or
- a birth weight greater than or equal to 1,500 grams (Birth Weight Category 6, 7, 8, or 9) and any-listed ICD-9-CM or ICD-10 PCS procedure codes for operating room procedure; or
- a birth weight greater than or equal to 1,500 grams (Birth Weight Category 6, 7, 8, or 9) and any-listed ICD-9-CM or ICD-10 PCS procedure codes for mechanical ventilation; or
- a birth weight greater than or equal to 1,500 grams (Birth Weight Category 6, 7, 8, or 9) and transferring from another health care facility within two days of birth

See Pediatric Quality Indicators Appendices:

- Appendix A – Operating Room Procedure Codes
- Appendix I – Definitions of Neonate, Newborn, Normal Newborn, and Outborn
- Appendix L – Low Birth Weight Categories

S.8. **Target Population Category** *(Check all the populations for which the measure is specified and tested if any):*
Children's Health

S.9. **Denominator Details** *(All information required to identify and calculate the target population/denominator such as definitions, specific data collection items/responses , code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
Please see attached excel file in S.2b. for version 6.1 alpha specifications.

S.10. **Denominator Exclusions** *(Brief narrative description of exclusions from the target population)*
Exclude cases:

- with a principal ICD-9-CM or ICD-10-CM diagnosis code (or secondary diagnosis present on admission†) for sepsis
- with a principal ICD-9-CM or ICD-10-CM diagnosis code (or secondary diagnosis present on admission†) for sepsis or bacteremia
- with a principal ICD-9-CM or ICD-10-CM diagnosis code (or secondary diagnosis present on admission†) for staphylococcal

or Gram-negative bacterial infection
- with birth weight less than 500 grams (Birth Weight Category 1)
- with length of stay less than 3 days
- with missing gender (SEX=missing), age (AGE=missing), quarter (DQTR=missing), year (YEAR=missing) or principal diagnosis (DX1=missing)

† Only for cases that otherwise qualify for the numerator.

S.11. **Denominator Exclusion Details** *(All information required to identify and calculate exclusions from the denominator such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
Please see attached excel file in S.2b. for version 6.1 alpha specifications.

S.12. **Stratification Details/Variables** *(All information required to stratify the measure results including the stratification variables, definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format with at S.2b)*
Not applicable

S.13. **Risk Adjustment Type** (Select type. Provide specifications for risk stratification in S.12 and for statistical model in S.14-15)
Statistical risk model
If other:

S.14. **Identify the statistical risk model method and variables** *(Name the statistical method - e.g., logistic regression and list all the risk factor variables. Note - risk model development and testing should be addressed with measure testing under Scientific Acceptability)*
The predicted value for each case is computed using a hierarchical model (logistic regression with hospital random effect) and covariates for gender, birthweight (in 500g groups), modified CMS DRG, congenital anomolies, transfer in status and Major Diagnosis Categories (MDC) based on principal diagnosis, availability of data elements (i.e., procedure days and point of origin). From these candidate variables, specific covariates were selected using a backward stepwise bootstrap approach. The expected rate is computed as the sum of the predicted value for each case divided by the number of cases for the unit of analysis of interest (i.e., hospital). The risk adjusted rate is computed using indirect standardization as the observed rate divided by the expected rate, multiplied by the reference population rate.
Additional information on methodology can be found in the Empirical Methods document on the AHRQ Quality Indicator website (www.qualityindicators.ahrq.gov). The Empirical Methods are also attached as "supplemental materials".

The specific covariates for this measure are as follows:

PARAMETER	LABEL
Intercept	Intercept
PD_MALE_1c	Male | Age Days = 0
PD_MALE_2c	Male | Age Days >= 1
PD_FEMALE_	Female | Age Days >= 1
BWHTCAT_1	Birth Weight 2500+ grams
BWHTCAT_2	Birth Weight 2000 to 2499 grams
BWHTCAT_3	Birth Weight 1750 to 1999 grams
BWHTCAT_4	Birth Weight 1500 to 1749 grams
BWHTCAT_5	Birth Weight 1250 to 1499 grams
BWHTCAT_6	Birth Weight 1000 to 1249 grams
BWHTCAT_7	Birth Weight 750 to 999 grams
BWHTCAT_8	Birth Weight <= 749 grams
BWHTCAT_1b	Birth Weight <= 749 grams
BWHTCAT_2b	Birth Weight > 749 grams
MDC_1	MDC 1: Nervous System
MDC_11	MDC 11: Kidney And Urinary Tract
MDC_15	MDC 15: Newborn And Other Neonates
MDC_3	MDC 3: Ear Nose Mouth And Throat
MDC_4	MDC 4: Respiratory System
MDC_5	MDC 5: Circulatory System
MDC_6	MDC 6: Digestive System

MDC_8	MDC 8: Musculoskeletal And Connective
NOPOUB04	No Point Of Origin
NOPRDAY	Days to procedure flag
mdrg_103	Craniotomy and endovascular intracranial procedure
mdrg_7799	ECMO or trach w MV 96+ hrs or PDX exc face
mdrg_8898	Ungroupable

Source: http://qualityindicators.ahrq.gov/Modules/pdi_resources.aspx
Parameter estimates are also included with the Technical Specifications attached in section S.2b
Please note Version 6.0 will be released publicly in March 2016.

S.15. **Detailed risk model specifications** *(must be in attached data dictionary/code list Excel or csv file. Also indicate if available at measure-specific URL identified in S.1.)*
Note: Risk model details (including coefficients, equations, codes with descriptors, definitions), should be provided on a separate worksheet in the suggested format in the Excel or csv file with data dictionary/code lists at S.2b.
Available in attached Excel or csv file at S.2b

S.15a. **Detailed risk model specifications** (*if not provided in excel or csv file at S.2b*)
Not applicable.

S.16. **Type of score:**
Rate/proportion
If other:

S.17. **Interpretation of Score** *(Classifies interpretation of score according to whether better quality is associated with a higher score, a lower score, a score falling within a defined interval, or a passing score)*
Better quality = Lower score

S.18. **Calculation Algorithm/Measure Logic** *(Describe the calculation of the measure score as an ordered sequence of steps including identifying the target population; exclusions; cases meeting the target process, condition, event, or outcome; aggregating data; risk adjustment; etc.)*
The observed rate is the number of discharge records where the patient experienced the QI adverse event divided by the number of discharge records at risk for the event. The expected rate is a comparative rate that incorporates information about a reference population that is not part of the user's input dataset – what rate would be observed if the expected level of care observed in the reference population and estimated with risk adjustment regression models, were applied to the mix of patients with demographic and comorbidity distributions observed in the user's dataset? The expected rate is calculated only for risk-adjusted indicators.

The expected rate is estimated for each person using a generalized estimating equations (GEE) approach to account for correlation at the hospital or provider level.

The risk-adjusted rate is a comparative rate that also incorporates information about a reference population that is not part of the input dataset – what rate would be observed if the level of care observed in the user's dataset were applied to a mix of patients with demographics and comorbidities distributed like the reference population? The risk adjusted rate is calculated using the indirect method as observed rate divided by expected rate multiplied by the reference population rate. The smoothed rate is the weighted average of the risk-adjusted rate from the user's input dataset and the rate observed in the reference population; the smoothed rate is calculated with a shrinkage estimator to result in a rate near that from the user's dataset if the provider's rate is estimated in a stable fashion with minimal noise, or to result in a rate near that of the reference population if the variance of the estimated rate from the input dataset is large compared with the hospital-to-hospital variance estimated from the reference population. Thus, the smoothed rate is a weighted average of the risk-adjusted rate and the reference population rate, where the weight is the signal-to-noise ratio. In practice, the smoothed rate brings rates toward the mean, and tends to do this more so for outliers (such as rural hospitals).

For additional information, please see the supplemental files for the Empirical Methods.

S.19. **Calculation Algorithm/Measure Logic Diagram URL or Attachment** *(You also may provide a diagram of the Calculation Algorithm/Measure Logic described above at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
No diagram provided

S.20. **Sampling** *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*
IF a PRO-PM, identify whether (and how) proxy responses are allowed.
Not applicable

S.21. **Survey/Patient-reported data** *(If measure is based on a survey, provide instructions for conducting the survey and guidance on minimum response rate.)*
IF a PRO-PM, specify calculation of response rates to be reported with performance measure results.
Not applicable

S.22. **Missing data** (specify how missing data are handled, e.g., imputation, delete case.)
Required for Composites and PRO-PMs.
Exclude cases with missing gender (SEX=missing), age (AGE=missing), quarter (DQTR=missing), year (YEAR=missing), or principal diagnosis (DX1=missing). Missingness on these variables, in aggregate, almost never exceeds 1% of eligible records.

S.23. **Data Source** *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*
If other, please describe in S.24.
Administrative claims

S.24. **Data Source or Collection Instrument** *(Identify the specific data source/data collection instrument e.g. name of database, clinical registry, collection instrument, etc.)*
IF a PRO-PM, identify the specific PROM(s); and standard methods, modes, and languages of administration.
While the measure is tested and specified using data from the Healthcare Cost and Utilization Project (HCUP) (see section 1.1 and 1.2 of the measure testing form), the measure specifications and software are specified to be used with any ICD-9-CM- or ICD-10-CM/PCS coded administrative billing/claims/discharge dataset.

S.25. **Data Source or Collection Instrument** *(available at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
Available at measure-specific web page URL identified in S.1

S.26. **Level of Analysis** *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*
Facility

S.27. **Care Setting** *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*
Hospital/Acute Care Facility
If other:

S.28. **COMPOSITE Performance Measure** - Additional Specifications *(Use this section as needed for aggregation and weighting rules, or calculation of individual performance measures if not individually endorsed.)*
Not applicable

2a. Reliability – See attached Measure Testing Submission Form
2b. Validity – See attached Measure Testing Submission Form
NQI03_Measure_Testing_Form_160212.docx

NATIONAL QUALITY FORUM—Measure Testing (subcriteria 2a2, 2b2-2b7)

Measure Number (*if previously endorsed*)**:** 0478
Measure Title: Neonatal Blood Stream Infection Rate (NQI 03)
Date of Submission: 2/16/2016
Type of Measure:

☐ Composite – ***STOP – use composite testing form***	☒ Outcome (*including PRO-PM*)
☐ Cost/resource	☐ Process
☐ Efficiency	☐ Structure

Instructions

- Measures must be tested for all the data sources and levels of analyses that are specified. ***If there is more than one set of data specifications or more than one level of analysis, contact NQF staff*** about how to present all the testing information in one form.
- **For all measures, sections 1, 2a2, 2b2, 2b3, and 2b5 must be completed.**
- **For outcome and resource use measures**, section **2b4** also must be completed.
- If specified for **multiple data sources/sets of specificaitons** (e.g., claims and EHRs), section **2b6** also must be completed.
- Respond to all questions as instructed with answers immediately following the question. All information on testing to demonstrate meeting the subcriteria for reliability (2a2) and validity (2b2-2b6) must be in this form. An appendix for *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 20 pages (*incuding questions/instructions;* minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.
- For information on the most updated guidance on how to address sociodemographic variables and testing in this form refer to the release notes for version 6.6 of the Measure Testing Attachment.

Note: The information provided in this form is intended to aid the Steering Committee and other stakeholders in understanding to what degree the testing results for this measure meet NQF's evaluation criteria for testing.

2a2. Reliability testing [10] demonstrates the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise. For **PRO-PMs and composite performance measures**, reliability should be demonstrated for the computed performance score.

2b2. Validity testing [11] demonstrates that the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality. For **PRO-PMs and composite performance measures**, validity should be demonstrated for the computed performance score.

2b3. Exclusions are supported by the clinical evidence; otherwise, they are supported by evidence of sufficient frequency of occurrence so that results are distorted without the exclusion; [12]

AND

If patient preference (e.g., informed decision making) is a basis for exclusion, there must be evidence that the exclusion impacts performance on the measure; in such cases, the measure must be specified so that the information about patient preference and the effect on the measure is transparent (e.g., numerator category computed separately, denominator exclusion category computed separately). [13]

2b4. For outcome measures and other measures when indicated (e.g., resource use):

- **an evidence-based risk-adjustment strategy** (e.g., risk models, risk stratification) is specified; is based on patient factors (including clinical and sociodemographic factors) that influence the measured outcome and are present at start of care; [14,15] and has demonstrated adequate discrimination and calibration

OR

- rationale/data support no risk adjustment/ stratification.

2b5. Data analysis of computed measure scores demonstrates that methods for scoring and analysis of the specified measure allow for **identification of statistically significant and practically/clinically meaningful** [16] **differences in performance**;

OR

there is evidence of overall less-than-optimal performance.

2b6. If multiple data sources/methods are specified, there is demonstration they produce comparable results.

2b7. For **eMeasures, composites, and PRO-PMs** (or other measures susceptible to missing data), analyses identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias.

Notes

10. Reliability testing applies to both the data elements and computed measure score. Examples of reliability testing for data elements include, but are not limited to: inter-rater/abstractor or intra-rater/abstractor studies; internal consistency for multi-item scales; test-retest for survey items. Reliability testing of the measure score addresses precision of measurement (e.g., signal-to-noise).

11. Validity testing applies to both the data elements and computed measure score. Validity testing of data elements typically analyzes agreement with another authoritative source of the same information. Examples of validity testing of the measure score include, but are not limited to: testing hypotheses that the measures scores indicate quality of care, e.g., measure scores are different for groups known to have differences in quality assessed by another valid quality measure or method; correlation of measure scores with another valid indicator of quality for the specific topic; or relationship to conceptually related measures (e.g., scores on process measures to scores on outcome measures). Face validity of the measure score as a quality indicator may be adequate if accomplished through a systematic and transparent process, by identified experts, and explicitly addresses whether performance scores resulting from the measure as specified can be used to distinguish good from poor quality.

12. Examples of evidence that an exclusion distorts measure results include, but are not limited to: frequency of occurrence, variability of exclusions across providers, and sensitivity analyses with and without the exclusion.

13. Patient preference is not a clinical exception to eligibility and can be influenced by provider interventions.

14. Risk factors that influence outcomes should not be specified as exclusions

15. With large enough sample sizes, small differences that are statistically significant may or may not be practically or clinically meaningful. The substantive question may be, for example, whether a statistically significant difference of one percentage point in

the percentage of patients who received smoking cessation counseling (e.g., 74 percent v. 75 percent) is clinically meaningful; or whether a statistically significant difference of $25 in cost for an episode of care (e.g., $5,000 v. $5,025) is practically meaningful. Measures with overall less-than-optimal performance may not demonstrate much variability across providers.

1. DATA/SAMPLE USED FOR ALL TESTING OF THIS MEASURE

Often the same data are used for all aspects of measure testing. In an effort to eliminate duplication, the first five questions apply to all measure testing. If there are differences by aspect of testing,(e.g., reliability vs. validity) be sure to indicate the specific differences in question 1.7.

1.1. What type of data was used for testing? (*Check all the sources of data identified in the measure specifications and data used for testing the measure. Testing must be provided for all the sources of data specified and intended for measure implementation.* ***If different data sources are used for the numerator and denominator, indicate N [numerator] or D [denominator] after the checkbox.***)

Measure Specified to Use Data From: (***must be consistent with data sources entered in S.23***)	**Measure Tested with Data From:**
☐ abstracted from paper record	☐ abstracted from paper record
☒ administrative claims	☒ administrative claims
☐ clinical database/registry	☐ clinical database/registry
☐ abstracted from electronic health record	☐ abstracted from electronic health record
☐ eMeasure (HQMF) implemented in EHRs	☐ eMeasure (HQMF) implemented in EHRs
☐ other: Click here to describe	☐ other: Click here to describe

1.2. If an existing dataset was used, identify the specific dataset (*the dataset used for testing must be consistent with the measure specifications for target population and healthcare entities being measured; e.g., Medicare Part A claims, Medicaid claims, other commercial insurance, nursing home MDS, home health OASIS, clinical registry*).

All analyses were completed using data from the Healthcare Cost and Utilization Project (HCUP) State Inpatient Databases (SID), 2011-2013. HCUP is a family of health care databases and related software tools and products developed through a Federal-State-Industry partnership and sponsored by the Agency for Healthcare Research and Quality (AHRQ).[1] HCUP databases bring together the data collection efforts of State data organizations, hospital associations, private data organizations, and the Federal government to create a national information resource of encounter-level health care data. The HCUP SID contain the universe of the inpatient discharge abstracts in participating States, translated into a uniform format to facilitate multi-State comparisons and analyses. All states provide data for community hospitals and together, the SID encompasses about 97 percent of all U.S. community hospital discharges. For the analyses presented here, we use 34 states representing about 89percent of the U.S. community hospital discharges, for a total of about 30 million hospital discharges from community hospitals. As defined by the American Hospital Association, community hospitals are all non-Federal, short-term, general or other specialty hospitals, excluding hospital units of institutions. Included among community hospitals are public and academic medical centers, specialty hospitals such as obstetrics–gynecology, ear–nose–throat, orthopedic and pediatric institutions. Short-stay

[1] HCUP State Inpatient Databases (SID). Healthcare Cost and Utilization Project (HCUP). 2011-2013. Agency for Healthcare Research and Quality, Rockville, MD. www.hcup-us.ahrq.gov/sidoverview.jsp. (AHRQ QI Software Version 6.1 alpha)

rehabilitation, long-term acute care hospitals are excluded from the data used for the reported analyses.

Each of the 34 states included in the dataset report information about whether a diagnosis was present on admission (POA) and information on the timing of procedures during the hospitalization. POA data[2] is important to distinguish complications that occur in-hospital from diagnoses that existed prior to hospitalization. Edit checks on POA were developed using a separate analysis of HCUP databases that examined POA coding in the 2013 SID at hospitals that were required to report POA to CMS. The edits identify general patterns of suspect reporting of POA. The edits do not evaluate whether a valid POA value (e.g., Y or N) is appropriate for the specific diagnosis. There are three hospital-level edit checks:

1. Indication that a hospital has POA reported as Y on all diagnoses on all discharges
2. Indication that a hospital has POA reported as missing on all non-Medicare discharges
3. Indication that a hospital reported POA as missing on all nonexempt diagnoses for 15 percent or more of discharges. The cut-point of 15 percent was determined by 2 times the standard deviation plus the mean of the percentage for hospitals required to report POA to CMS.

Hospitals that failed any of the edit checks were excluded from the dataset.

The SID data elements include ICD-9-CM coded principal and secondary diagnoses and procedures, additional detailed clinical and service information based on revenue codes, admission and discharge status, patient demographics, expected payment source (Medicare, Medicaid, private insurance as well as the uninsured), total charges and length of stay (www.hcup-us.ahrq.gov).

1.3. What are the dates of the data used in testing?
HCUP data included 2011-2013

1.4. What levels of analysis were tested? (*testing must be provided for <u>all</u> the levels specified and intended for measure implementation, e.g., individual clinician, hospital, health plan*)

Measure Specified to Measure Performance of: (***must be consistent with levels entered in item S.26***)	**Measure Tested at Level of:**
☐ individual clinician	☐ individual clinician
☐ group/practice	☐ group/practice
☒ hospital/facility/agency	☒ hospital/facility/agency
☐ health plan	☐ health plan
☐ other: Click here to describe	☐ other: Click here to describe

1.5. How many and which <u>measured entities</u> were included in the testing and analysis (by level of analysis and data source)? (*identify the number and descriptive characteristics of measured entities included in the analysis (e.g., size, location, type); if a sample was used, describe how entities were selected for inclusion in the sample*)

[2] Present-on -Admission was added as a data element to the uniform bill form (UB-04) effective October 1, 2007, and hospitals incurred a payment penalty for not including POA on Medicare records beginning October 1, 2008. Each of the several diagnoses in a discharge record can be flagged as "present at the time the order for inpatient admission occurs" or not (see http://www.cdc.gov/nchs/icd/icd9cm_addenda_guidelines.htm).

Table 1. Reference Population Rate and Distribution of Hospital Performance for NQI 03 Neonatal Blood Stream Infection Rate

Overall Reference Population Rate				
Year[3]	Number of Hospitals	Outcome of Interest (Numerator)[1]	Population at Risk (Denominator)[1]	Observed Rate Per 1000[1]
2011	1,285	1,746	72,697	24.018
2012	1,344	1,695	74,032	22.896
2013	1,277	1,331	68,647	19.389

Distribution of Hospital-level Observed Rates in Reference Population Per 1000								
Year[3]	Number of Hospitals	(p=percentile)[2]						
		Mean	SD	p5	p25	Median	p75	p95
2011	1,285	11.53	22.62	0.00	0.00	0.00	16.81	56.14
2012	1,344	11.62	26.61	0.00	0.00	0.00	14.04	58.82
2013	1,277	9.15	19.05	0.00	0.00	0.00	11.83	49.28

Source: HCUP State Inpatient Databases (SID). Healthcare Cost and Utilization Project (HCUP). 2011-2013. Agency for Healthcare Research and Quality, Rockville, MD. www.hcup-us.ahrq.gov/sidoverview.jsp. (AHRQ QI Software Version 6.1 alpha)

[1]The observed rate refers to the total rate for all observations included in the reference population data (numerator) divided by the total combined eligible population of all hospitals included in the reference population data (denominator).

[2]The distribution of hospital rates reports the mean and standard deviation (SD) of the observed rates for all hospitals included in the dataset, as well as the observed rate for hospitals in the 5th, 25th, 50th (median), 75th, and 95th percentile.

[3] Reference population is limited to states with present on admission data (POA). Since many states did not report POA data prior to 2011 we have not included testing prior to 2011.

1.6. How many and which <u>patients</u> were included in the testing and analysis (by level of analysis and data source)? (*identify the number and descriptive characteristics of patients included in the analysis (e.g., age, sex, race, diagnosis); if a sample was used, describe how patients were selected for inclusion in the sample*)
See 1.5 (Table 1)

1.7. If there are differences in the data or sample used for different aspects of testing (e.g., reliability, validity, exclusions, risk adjustment), identify how the data or sample are different for each aspect of testing reported below.
Some tests require comparisons of three years of data (2011-2013). When no comparisons are required for the test, only 2013 data are used.

1.8 What were the patient-level sociodemographic (SDS) variables that were available and analyzed in the data or sample used? For example, patient-reported data (e.g., income, education, language), proxy variables when SDS data are not collected from each patient (e.g. census tract), or patient community characteristics (e.g. percent vacant housing, crime rate).

Age and sex were the only patient-level sociodemographic variables that were available and analyzed in the data used for measure development and testing. The development data sets generally include race/ethnicity, principal expected source of payment, and zip code of residence, which could be used to capture socioeconomic characteristics at an ecological (community) level. However, these variables were not used in our analyses, based on our conceptual description (logical rationale or theory informed by literature and content experts) of the causal pathway between these factors, patient clinical factors, quality of care, and outcome, described in Section 2b4.3 below.

2a2. RELIABILITY TESTING

***Note**: If accuracy/correctness (validity) of data elements was empirically tested, separate reliability testing of data elements is not required – in 2a2.1 check critical data elements; in 2a2.2 enter "see section 2b2 for validity testing of data elements"; and skip 2a2.3 and 2a2.4.*

2a2.1. What level of reliability testing was conducted? (*may be one or both levels*)

☐ **Critical data elements used in the measure** (*e.g., inter-abstractor reliability; data element reliability must address ALL critical data elements*)

☒ **Performance measure score** (e.g., *signal-to-noise analysis*)

2a2.2. For each level checked above, describe the method of reliability testing and what it tests (*describe the steps—do not just name a method; what type of error does it test; what statistical analysis was used*)

The following text is outdated; a more complete description is immediately following: The signal to noise ratio is the ratio of the between hospital variance (signal) to the within hospital variance (noise). The formula is signal / (signal + noise). The ratio itself is only a diagnostic for the degree of variance in the risk-adjusted rate systematically associated with the provider. Therefore, what matters is the magnitude of the variance in the "smoothed" rate (that is, the variance in the risk-adjusted rate after the application of the univariate shrinkage estimator based on the signal ratio).

Signal to Noise. The signal-to-noise ratio refers to the entire population of US hospitals, comparing the degree to which risk adjusted rates are different from hospital to hospital (the signal) to how precise the rates are within hospitals (the noise). This metric is a stringent measure of reliability that takes into account the observed distribution of risk adjusted rates within a reference population. An indicator with a low signal-to-noise ratio may not be able to distinguish differences in performance between hospitals, or may identify differences inconsistently within the same time period. An indicator with a high signal-to-noise ratio will be more likely to consistently distinguish performance differences between hospitals (e.g. one hospital performs better than others).

The signal-to-noise ratio is estimated for each hospital. The overall signal-to-noise estimate is an average of hospital-level signal to noise ratios weighted by a value of one divided by the signal plus the hospital's noise for NQI 03. Hospitals with smaller denominators (the number of patients at risk) will have lower weight, and less influence on the overall signal-to-noise ratio, because of higher noise. Weighting reduces the influence of hospitals that have less reliable rates due to very small denominators (the number of patients at risk) on the overall signal-to-noise ratio estimate.

Because the signal-to-noise ratio quantifies the ability to consistently discriminate one hospital's performance from the other hospitals in the population, it is sensitive to the distribution of hospital sizes as well as the distribution of risk-adjusted rates in the reference population. If the hospitals in a population all have performance in a narrow range (low signal), it is more difficult to reliably distinguish between hospitals' performance than when hospital performance is spread out over a much wider range (high signal). For example, if all hospitals have nearly perfect performance, it will be impossible to distinguish between them. As a consequence, if the distribution of hospital rates changes over time, the signal-to-noise ratio will also change.

There is no universally accepted threshold of "adequate" signal to noise ratio. Different methods of calculating reliability and signal-to-noise result in different distributions of reliability scores. In addition, "adequate" depends on the specific application and judgment of the user. For instance, if a complication such as mortality is very important (e.g. leads to great harm to the patient) a lower reliability may be acceptable. However, the AHRQ QI program generally considers ratios between 0.4 – 0.8 as acceptable. It is rare to achieve reliability above 0.8. To account for the uncertainty (noise) in a hospital's performance due to reliability concerns stemming from low volume, smoothed rates can be calculated.

2a2.3. For each level of testing checked above, what were the statistical results from reliability testing? (e.g., *percent agreement and kappa for the critical data elements; distribution of reliability statistics from a signal-to-noise analysis*)

The following text is outdated: Updated Testing Results including both benign and malignant cases:
What the data demonstrate is systematic variation in the provider level rate of 0.419 to 69.167 per 1,000 from the 5th to 95th percentile after a signal ratio of 0.831 is applied as the shrinkage estimator (that is, after accounting for variation due to random factors).

HCUP State Inpatient Databases (SID). Healthcare Cost and Utilization Project (HCUP). 2008. Agency for Healthcare Research and Quality, Rockville, MD. Includes approximately 6 million pediatric discharges for 2,500 hospitals.

Table 2 shows the most recent reliability testing for NQI 03.

Table 2. Signal-to-Noise Ratio by Size Decile for NQI 03 Neonatal Blood Stream Infection Rate

Size Decile	Number of Hospitals	Avg. Number of Discharges per Hospital in Decile	Avg. Signal-to-Noise Ratio for Hospitals in Decile
1	94	3.5	0.03776
2	94	6.5	0.04589
3	95	12.4	0.12740
4	94	20.0	0.19443
5	94	31.6	0.31406
6	95	48.0	0.40531
7	94	68.4	0.47076
8	95	101.3	0.58954
9	94	150.3	0.69027
10	94	281.9	0.79322
Overall	943	72.3	0.63456

Source: HCUP State Inpatient Databases (SID). Healthcare Cost and Utilization Project (HCUP). 2013. Agency for Healthcare Research and Quality, Rockville, MD. www.hcup-us.ahrq.gov/sidoverview.jsp. (AHRQ QI Software Version 6.1 alpha)

2a2.4 What is your interpretation of the results in terms of demonstrating reliability? (i.e., *what do the results mean and what are the norms for the test conducted?*)
Signal to noise ratios were smaller for hospitals with fewer than 32 qualifying discharges per year (average signal-to-noise ratio less than 0.31). Smoothed rates, which are recommended for all hospitals (and are implemented in the AHRQ software), address reliability concerns particularly for small hospitals. Hospitals with more than 48 discharges on average have risk adjusted rates with moderate to high reliability (average signal-to-noise ratio of 0.41 to 0.79). Overall, the signal to noise ratio for this indicator is good with an overall signal-to-noise ratio of 0.63.

2b2. VALIDITY TESTING

2b2.1. What level of validity testing was conducted? (*may be one or both levels*)

☐ **Critical data elements** (*data element validity must address ALL critical data elements*)

☒ **Performance measure score**

☐ **Empirical validity testing**

☒ **Systematic assessment of face validity of performance measure score as an indicator** of quality or resource use (*i.e., is an accurate reflection of performance on quality or resource use and can distinguish good from poor performance*)

2b2.2. For each level of testing checked above, describe the method of validity testing and what it tests (*describe the steps—do not just name a method; what was tested, e.g., accuracy of data elements compared to authoritative source, relationship to another measure as expected; what statistical analysis was used)*

A structured panel review of each indicator was undertaken to evaluate the face validity (from a clinical perspective) of the indicator. Specifically, the panels approach sought to establish consensual validity, which "extends face validity from one expert to a panel of experts who examine and rate the appropriateness of each item…." The methodology for the structured review was adapted from the RAND/UCLA Appropriateness Method and consisted of an initial independent assessment of each indicator by clinician panelists using an initial questionnaire, a conference call among all panelists, followed by a final independent assessment by clinician panelists using the same questionnaire. The panel process served to refine definitions of some indicators, add new measures, and dismiss indicators with major concerns from further consideration.

Measures Of Pediatric Health Care Quality Based On Hospital Administrative Data
The Pediatric Quality Indicators Neonatal Indicator Appendix (2008)
http://qualityindicators.ahrq.gov/Downloads/Modules_Non_Software/Modules%20Development%20Bullet/pdi_development.zip

2b2.3. What were the statistical results from validity testing? (*e.g., correlation; t-test*)

The multi-specialty Panel rated the indicator as acceptable on overall usefulness as an indicator of potentially preventable complications of care.

Overall panel rating on a scale from 1-9

Question	**Median**	**Agreement status**
Overall rating of usefulness for internal QI improvement	8	Agree
Overall rating of usefulness for comparative purposes	8	Agree
Importance	8.5	Agree
Preventability	7	Indeterminate
Likelihood of medical error	6.5	Indeterminate
Charting	8	Agree
Bias	5	Indeterminate
Final recommendation		
Internal Quality Improvement: Recommend		
Comparative purposes: Recommend		

HCUP Kids' Inpatient Database (KID). Healthcare Cost and Utilization Project (HCUP). 2000 and 2003. Agency for Healthcare Research and Quality, Rockville, MD. www.hcup-us.ahrq.gov/kidoverview.jsp

2b2.4. What is your interpretation of the results in terms of demonstrating validity? (i.*e., what do the results mean and what are the norms for the test conducted?*)
The indicator has high face validity for use in quality improvement and hospital comparative purposes.

2b3. EXCLUSIONS ANALYSIS
NA ☐ no exclusions — *skip to section 2b4*

2b3.1. Describe the method of testing exclusions and what it tests (*describe the steps—do not just name a method; what was tested, e.g., whether exclusions affect overall performance scores; what statistical analysis was used*)

Clinical review of Exclusions: Clinical panel review (see above for a description of the methodology)
Exclusions were intended to identify cases where the outcome of interest was more likely to be present on admission.

Empirical Evaluation of Exclusions: Using the 2013 data from 34 states, we examined the percent of potential denominator cases excluded by each criterion as listed in the measure specifications.

2b3.2. What were the statistical results from testing exclusions? (*include overall number and percentage of individuals excluded, frequency distribution of exclusions across measured entities, and impact on performance measure scores*)

Clinical review of Exclusions: Changes Implemented as a Result of Pediatric Panel Review
Admissions from home or late transfers excluded. Patients admitted from home may have acquired the infection at home. Likewise, patients transferred on or after day two of age, may have acquired the infection at the transferring facility.
NOTE: In the current specification this is implemented as an inclusion criteria
- age in days less than 2 AND transferred from another health care facility

Exclude patients with a length of stay of less than 2 days - It is unlikely that these patients would acquire a nosocomial pathogen in such a short timespan.

HCUP Kids' Inpatient Database (KID). Healthcare Cost and Utilization Project (HCUP). 2000 and 2003. Agency for Healthcare Research and Quality, Rockville, MD. www.hcup-us.ahrq.gov/kidoverview.jsp

Empirical Evaluation of Exclusions: Table 3 shows the results of the most recent exclusions analysis.

Table 3. Number and Percent of Discharges Excluded, by Denominator Exclusion Criteria for NQI 03 Neonatal Blood Stream Infection Rate[1]

NQI 03	Denominator			Potential Numerator[2]		
`Exclusion Name	Exclusion Count	After Exclusions	% Change	Exclusion Count	After Exclusions	% Change
Exclude Principal Diagnosis of Sepsis	350	2,988,719	0.01%	3	2,832	0.11%
Exclude birthweight < 500g	725	2,988,344	0.02%	50	2,785	1.76%
Exclude LOS < 3 days	19,967	2,969,102	0.67%	43	2,792	1.52%
No Exclusions applied	-	**2,989,069**	-	-	**2,835**	-

[1]This indicator does not have numerator exclusion criteria.
[2]Potential numerator cases are those that would have qualified for the numerator if not for a particular denominator exclusion criterion.
Source: HCUP State Inpatient Databases (SID). Healthcare Cost and Utilization Project (HCUP). 2013. Agency for Healthcare Research and Quality, Rockville, MD. www.hcup-us.ahrq.gov/sidoverview.jsp. (AHRQ QI Software Version 6.1 alpha)

2b3.3. What is your interpretation of the results in terms of demonstrating that exclusions are needed to prevent unfair distortion of performance results? (*i.e., the value outweighs the burden of increased data collection and analysis. Note: **If patient preference is an exclusion**, the measure must be specified so that the effect on the performance score is transparent, e.g., scores with and without exclusion*)

Patients with a principal diagnosis of sepsis are excluded because these infections are present on admission. Patients with a LOS of less than 3 days are excluded because sepsis cases in babies with a shorter LOS are likely to be perinatally acquired, and the transfer out rate is highest in the first 2 days of life. Patients with this short of a LOS have low risk of exposure to postnatally acquired sepsis.
Patients with a birthweight of less than 500g are excluded from all NQI because these babies are very fragile and high risk; sepsis and other poor outcomes are less preventable in this population. Overall the exclusions account for less than 1% of denominator cases and less than 2% of numerator cases. Exclusions are retained to ensure face validity of the measure. Since many hospitals have only a few cases that qualify for the numerator, retaining even rare exclusions will ensure fair comparisons between providers.

2b4. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES
If not an intermediate or health outcome, or PRO-PM, or resource use measure, skip to section 2b5.

2b4.1. What method of controlling for differences in case mix is used?
☐ **No risk adjustment or stratification**
☒ **Statistical risk model with 26 risk factors**
☐ **Stratification by** Click here to enter number of categories **risk categories**
☐ **Other,** Click here to enter description

2b4.2. If an outcome or resource use measure is not risk adjusted or stratified, provide rationale and analyses to demonstrate that controlling for differences in patient characteristics (case mix) is not needed to achieve fair comparisons across measured entities.
Not applicable

2b4.3. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors (clinical factors or sociodemographic factors) used in the statistical risk model or for stratification by risk (*e.g., potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of p<0.10; correlation of x or higher; patient factors should be present at the start of care*)

The following text is outdated: Risk-adjustment models use a standard set of categories based on readily available classification systems for demographics, severity of illness and comorbidities. Within each category, covariates are initially selected based on a minimum of 30 cases in the outcome of interest. Then a stepwise regression process on a development sample is used to select a parsimonious set of covariates where p<.05. Model is then tested on a validation sample.

The following text is outdated: If the user´s data lacks present on admission information, then the likelihood that the outcome of interest and the covariates are present on admission is estimated using a Markov Chain Monte Carlo (MCMC) estimation procedure. That likelihood is then used to adjust the observed and expected rates.

Clinical Factors
For the provider level indicators, each module has a standard set of covariates grouped into four categories: demographics, severity of illness, comorbidities and transfer-in status. The standard set is tailored to each indicator to create a parsimonious set of covariates for each indicator. Based on cross tabulations between each covariate and the outcome of interest, only those covariates with at least 30 cases with the outcome of interest are retained. For

categories that are mutually exclusive, covariates with fewer than 30 cases are pooled into the next covariate along the risk gradient. For categories with no risk gradient, covariates are pooled into broader covariates. Covariates that are considered as potential risk adjusters for the Neonatal Quality Indicators include age in days at admission, gender, birthweight, Major Diagnostic Categories (MDC), MS-DRG (removing comorbidity and complication distinctions), availability of data elements (point of origin and procedure dates) and whether they were transferred from another facility.

The omitted covariate within mutually exclusive categories is the reference group for those categories. Reference categories are usually 1) the most common and/or 2) the least risk.

The choice of omitted reference category does affect how one might use the model coefficients or odds ratios in an English language sentence, but it does not affect predicted probabilities or model performance.

Once the preliminary multivariable model is specified, it is estimated on the pediatric analytic data, as appropriate using a backward stepwise bootstrap approach.

Additional details are available in the AHRQ Quality Indicator Empirical Methods document, included in the supplemental materials.

Sociodemographic (SDS) Factors

Few studies have explored disparities in neonatal blood stream infection. Stoll et al (2002) reported that after adjusting for gestational age, birth weight, and clinical center, there was no relationship between either race or sex and the risk of late-onset sepsis. There is no evidence or causal model to suggest that SDS factors are associated with late onset sepsis independent of quality of care, or are mediated by pre-hospital care (which may not fall within the proper realm of hospital accountability. Accordingly, consistent with the guidance provided by NQF in the SDS Trial Period FAQs, AHRQ believes that it would be inappropriate to include other SDS variables in the risk-adjustment approach for NQI 03, which is an in-hospital outcome measure.

Stoll BJ, Hansen N, Fanaroff AA, et al. Late-onset sepsis in very low birth weight neonates: the experience of the NICHD Neonatal Research Network. *Pediatrics.* 2002;110(2 Pt 1):285-291.

2b4.4a. What were the statistical results of the analyses used to select risk factors?

Table 4 lists the current risk adjustment coefficients for NQI 03.

Table 4. Risk Adjustment Coefficients for NQI 03 Neonatal Blood Stream Infection Rate

Parameter	Label	DF	Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square
Intercept	Intercept	1	-3.2646	0.3193	104.5576	<.0001
PD_MALE_1c	Male | Age Days = 0	1	0.1484	0.0544	7.4351	0.0064
PD_MALE_2c	Male | Age Days >= 1	1	0.0581	0.1694	0.1175	0.7317
PD_FEMALE_	Female | Age Days >= 1	1	0.0594	0.1963	0.0917	0.762
BWHTCAT_1	Birth Weight 2500+ grams	1	-0.2312	0.1769	1.7088	0.1911
BWHTCAT_2	Birth Weight 2000 to 2499 grams	1	-0.0139	0.1546	0.0081	0.9283
BWHTCAT_3	Birth Weight 1750 to 1999 grams	1	0.3669	0.1722	4.5401	0.0331

BWHTCAT_4	Birth Weight 1500 to 1749 grams	1	0.3944	0.1525	6.6837	0.0097
BWHTCAT_5	Birth Weight 1250 to 1499 grams	1	0.3083	0.1241	6.1738	0.013
BWHTCAT_6	Birth Weight 1000 to 1249 grams	1	1.2207	0.1078	128.1888	<.0001
BWHTCAT_7	Birth Weight 750 to 999 grams	1	1.9086	0.1003	362.023	<.0001
BWHTCAT_8	Birth Weight <= 749 grams	1	2.3741	0.1003	560.461	<.0001
BWHTCAT_1b	Birth Weight <= 749 grams	0	0	0	.	<.0001
BWHTCAT_2b	Birth Weight > 749 grams	0	0	0	.	<.0001

Parameter	Label	DF	Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square
MDC_1	MDC 1: Nervous System	1	-2.6364	1.0481	6.3267	0.0119
MDC_11	MDC 11: Kidney And Urinary Tract	1	-1.3073	1.0534	1.5402	0.2146
MDC_15	MDC 15: Newborn And Other Neonates	1	-1.669	0.3151	28.0491	<.0001
MDC_3	MDC 3: Ear Nose Mouth And Throat	1	-0.8865	0.7867	1.2698	0.2598
MDC_4	MDC 4: Respiratory System	1	-0.9382	0.4602	4.1572	0.0415
MDC_5	MDC 5: Circulatory System	1	-0.8543	0.3104	7.573	0.0059
MDC_6	MDC 6: Digestive System	1	-1.015	0.3717	7.455	0.0063
MDC_8	MDC 8: Musculoskeletal And Connective	1	-0.8028	0.5857	1.8789	0.1705
NOPOUB04	No Point Of Origin	1	-0.1471	0.0803	3.3556	0.067
NOPRDAY	Days to procedure flag	1	-0.8909	0.1928	21.3446	<.0001
mdrg_103	Craniotomy and endovascular intracranial procedure	1	1.4721	1.4209	1.0733	0.3002
mdrg_7799	ECMO or trach w MV 96+ hrs or PDX exc face	1	1.427	0.1394	104.7977	<.0001
mdrg_8898	Ungroupable	1	0.5186	0.2665	3.7882	0.0516
c-statistic=0.752						

Source: HCUP State Inpatient Databases (SID). Healthcare Cost and Utilization Project (HCUP). 2013. Agency for Healthcare Research and Quality, Rockville, MD. www.hcup-us.ahrq.gov/sidoverview.jsp. (AHRQ QI Software Version 6.1 alpha)

2b4.4b. Describe the analyses and interpretation resulting in the decision to select SDS factors (e.g. prevalence of the factor across measured entities, empirical association with the outcome, contribution of unique variation in the outcome, assessment of between-unit effects and within-unit effects)
Not applicable (see above)

2b4.5. Describe the method of testing/analysis used to develop and validate the adequacy of the statistical model or stratification approach (*describe the steps—do not just name a method; what statistical analysis was used*)
This analysis evaluates how strongly the risk adjustment model is associated with the event of interest (i.e. hospital acquired neonatal blood stream infection). The measure of discrimination, how well the risk adjustment model distinguishes events from non-events, is the c-statistic. The c-statistic is computed by assigning each observation a predicted probability of the outcome from the risk-adjustment model based on the value of the observations covariates from the risk-adjustment model. Two copies of the dataset are sorted, first from highest to lowest predicted probability and second from lowest to highest predicted probability. This creates a set of pairs of observations. Pairs that consist of one event and one non-event (discordant pairs) are kept and concordant pairs are discarded. The c-statistic is a measure of the proportion of discordant pairs of observations for which the observation with the event had a higher predicted probability from the risk-adjustment model than the non-event. C-statistics above 0.70 and below 0.80 have moderate discrimination. Above 0.80 the discrimination is high. We did not employ common "goodness of fit" tests because these tests tend to not be informative with large samples.

We also evaluated the calibration of the risk adjustment model by evaluating how closely observed and predicted rates compare across deciles of the predicted rate. This analysis splits the sample into deciles based on predicted rates, and then compares these rates with the observed rates for the population in each decile. A well calibrated model, or one that does not over or under-estimate risk, will have comparable observed and predicted rates across the risk spectrum.

Provide the statistical results from testing the approach to controlling for differences in patient characteristics (case mix) below.
If stratified, skip to 2b4.9

2b4.6. Statistical Risk Model Discrimination Statistics (*e.g., c-statistic, R-squared*)**:**
The following text is outdated:
c-statistic for the outcome of interest (y|x): 0.614
c-statistic for present on admission (p|x): 0.718 (when POA data are missing)

HCUP State Inpatient Databases (SID). Healthcare Cost and Utilization Project (HCUP). 2008. Agency for Healthcare Research and Quality, Rockville, MD. Includes approximately 6 million pediatric discharges for 2,500 hospitals.

Table 5 lists the current results for risk adjustment model performance.

Table 5. Risk adjustment Model Discrimination and Calibration, for NQI 03 Neonatal Blood Stream Infection Rate

Predicted Rate Decile	Number of Discharges per Decile	Predicted Rate	Observed Rate
1	6,864	5.48800	4.66200
2	6,865	7.06900	6.70100
3	6,865	7.93500	5.82700
4	6,865	8.58200	5.53500
5	6,864	9.96800	10.05200

Predicted Rate Decile	Number of Discharges per Decile	Predicted Rate	Observed Rate
6	6,865	11.19700	11.79900
7	6,865	14.90200	15.14900
8	6,865	25.36200	23.88900
9	6,865	42.28300	39.33000
10	6,864	72.32700	70.95000
C-Statistic 0.752			

Source: HCUP State Inpatient Databases (SID). Healthcare Cost and Utilization Project (HCUP). 2013. Agency for Healthcare Research and Quality, Rockville, MD. www.hcup-us.ahrq.gov/sidoverview.jsp. (AHRQ QI Software Version 6.1 alpha)

2b4.7. Statistical Risk Model Calibration Statistics (*e.g., Hosmer-Lemeshow statistic*):
See Table 5 in 2b4.6

2b4.8. Statistical Risk Model Calibration – Risk decile plots or calibration curves:
See Table 5 in 2b4.6

2b4.9. Results of Risk Stratification Analysis:
Not applicable

2b4.10. What is your interpretation of the results in terms of demonstrating adequacy of controlling for differences in patient characteristics (case mix)? (i.*e., what do the results mean and what are the norms for the test conducted*)
A model that is well calibrated will have observed values similar to predicted values across the predicted value deciles. This indicator is well calibrated and has good discrimination, as the observed to predicted values across the deciles range between 0.64 – 1.05. The ratio falls below 0.85 for only 2 deciles. The C-statistic is good at 0.75.

2b4.11. Optional Additional Testing for Risk Adjustment (*not required, but would provide additional support of adequacy of risk model, e.g., testing of risk model in another data set; sensitivity analysis for missing data; other methods that were assessed*)

2b5. IDENTIFICATION OF STATISTICALLY SIGNIFICANT & MEANINGFUL DIFFERENCES IN PERFORMANCE

2b5.1. Describe the method for determining if statistically significant and clinically/practically meaningful differences in performance measure scores among the measured entities can be identified (*describe the steps—do not just name a method; what statistical analysis was used? Do not just repeat the information provided related to performance gap in 1b)*
The following text is outdated:
Posterior probability distribution parameterized using the Gamma distribution

Current methods:
This analysis assesses the probability that a hospital is higher or lower than a benchmark or threshold, given hospital size. It reflects whether the indicator can discriminate the best performing hospitals from the lower performing hospitals.

For this analysis, "benchmark" refers to the smoothed indicator rate based on the 20th percentile of the reference population (i.e., 20% of hospitals have a lower mortality rate or better performance). "Threshold" refers to the indicator rate based on the 80th percentile (i.e., 80% have lower mortality or better performance).

The analysis is reported by size decile, based on the denominator cases, demonstrating performance across hospitals of various sizes. Each hospital is assumed to have an underlying distribution of smoothed rates that follows a Gamma distribution. The parameters of a Gamma distribution are shape and scale. For each hospital the shape is calculated as ((smoothed rate)^2/ smoothed rate variance), and the scale is calculated as (smoothed rate variance / smoothed rate). The smoothed rate variance (aka posterior variance) is calculated as the signal variance – (reliability weight * signal variance). The reliability weight is calculated as (signal variance / (signal variance + noise variance)). Hospitals are ranked by size and grouped into 10 equal categories of size (deciles). The Benchmark and Threshold are compared to the Gamma distribution of the smoothed rates for each hospital to determine if the hospital rate is better or worse than the Benchmark and Threshold rates with 95% probability. This provides a 95% confidence interval for the Benchmark and Threshold rate.

Table 6 reports the proportion of hospitals above (better than) and below (worse than) the Benchmark and Threshold rates and the proportion not classified as either above or below. The proportion of hospitals not classified as either better or worse have rates that fall within the 95% confidence interval.

2b5.2. What were the statistical results from testing the ability to identify statistically significant and/or clinically/practically meaningful differences in performance measure scores across measured entities? (e.g., *number and percentage of entities with scores that were statistically significantly different from mean or some benchmark, different from expected; how was meaningful difference defined*)

The following text is outdated:

Raw Rates (numerator / denominator):

5th 25th Median 75th 95th

0.000419 0.004075 0.012324 0.028426 0.069167

HCUP State Inpatient Databases (SID). Healthcare Cost and Utilization Project (HCUP). 2008. Agency for Healthcare Research and Quality, Rockville, MD. Includes approximately 6 million pediatric discharges for 2,500 hospitals.

Table 6 shows the current performance discrimination results for NQI 03.

Table 6. Performance Categories by Hospital Size Decile for NQI 03 Neonatal Blood Stream Infection Rate

			Benchmark			Threshold		
Size Decile	Number of Hospitals	Average Number of Denominator Population Per Hospital	Proportion Better	Proportion Worse	Proportion Unclassified	Proportion Better	Proportion Worse	Proportion Unclassified
1	94	3.5	0.0000	0.0000	1.0000	0.0000	0.0000	1.0000
2	94	6.5	0.0000	0.0000	1.0000	0.0000	0.0000	1.0000
3	95	12.4	0.0000	0.0421	0.9579	0.0000	0.0000	1.0000
4	94	20	0.0000	0.1064	0.8936	0.0000	0.0000	1.0000
5	94	31.6	0.0000	0.1383	0.8617	0.0000	0.0000	1.0000
6	95	48	0.0000	0.2632	0.7368	0.0105	0.0421	0.9474
7	94	68.4	0.0000	0.2979	0.7021	0.0319	0.0106	0.9575
8	95	101.3	0.0000	0.2947	0.7053	0.1789	0.0105	0.8106
9	94	150.3	0.0000	0.5426	0.4574	0.2979	0.0532	0.6489
10	94	281.9	0.0000	0.5319	0.4681	0.4787	0.0532	0.4681
Overall	943	72.3	0.0000	0.2216	0.7784	0.0997	0.0170	0.8833

Source: HCUP State Inpatient Databases (SID). Healthcare Cost and Utilization Project (HCUP). 2013. Agency for Healthcare Research and Quality, Rockville, MD. www.hcup-us.ahrq.gov/sidoverview.jsp. (AHRQ QI Software Version 6.1 alpha)

2b5.3. What is your interpretation of the results in terms of demonstrating the ability to identify statistically significant and/or clinically/practically meaningful differences in performance across measured entities? (i.*e., what do the results mean in terms of statistical and meaningful differences?*)

This indicator has poor discrimination for most hospitals to identify low performing hospitals; 11% of hospitals can be classified as better or worse than the threshold (the percentage classified as either above or below the threshold). The indicator has poor discrimination, particularly for moderate to large hospitals to identify high performing hospitals; 22% of hospitals can be classified as better or worse than the benchmark. These results reflect the skewed distribution for this indicator, due to many hospitals with very small denominators and zero numerators. For the largest hospitals with 159 or more high risk infants in the denominator, performance discrimination improves with about 55% of hospitals classified as better or worse than the benchmark and 35% - 47% hospitals classified as better or worse than the threshold.

2b6. COMPARABILITY OF PERFORMANCE SCORES WHEN MORE THAN ONE SET OF SPECIFICATIONS
If only one set of specifications, this section can be skipped.

Note: *This item is directed to measures that are risk-adjusted (with or without SDS factors)* ***OR*** *to measures with more than one set of specifications/instructions (e.g., one set of specifications for how to identify and compute the measure from medical record abstraction and a different set of specifications for claims or eMeasures). It does not apply to measures that use more than one source of data in one set of specifications/instructions (e.g., claims data to identify the denominator and medical record abstraction for the numerator).* ***Comparability is not required when comparing performance scores with and without SDS factors in the risk adjustment model. However, if comparability is not demonstrated for measures with more than one set of specifications/instructions, the different specifications (e.g., for medical records vs. claims) should be submitted as separate measures.***

Not applicable

2b6.1. Describe the method of testing conducted to compare performance scores for the same entities across the different data sources/specifications (*describe the steps—do not just name a method; what statistical analysis was used*)
Not applicable

2b6.2. What were the statistical results from testing comparability of performance scores for the same entities when using different data sources/specifications? (*e.g., correlation, rank order*)
Not applicable

2b6.3. What is your interpretation of the results in terms of the differences in performance measure scores for the same entities across the different data sources/specifications? (i.*e., what do the results mean and what are the norms for the test conducted*)
Not applicable

2b7. MISSING DATA ANALYSIS AND MINIMIZING BIAS

2b7.1. Describe the method of testing conducted to identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias (*describe the steps—do not just name a method; what statistical analysis was used*)

The AHRQ QIs use frequently reported administrative data variables. NQI 03 excludes cases with missing age, sex, discharge quarter, discharge year, and principal diagnosis. These variables are required for indicator construction and are required of all hospital discharge records. The rate of missing data for each variable is available by state and year from the AHRQ HCUP website (http://www.hcup-us.ahrq.gov/cdstats/cdstats_search.jsp).

2b7.2. What is the overall frequency of missing data, the distribution of missing data across providers, and the results from testing related to missing data? (*e.g., results of sensitivity analysis of the effect of various rules for missing data/nonresponse; if no empirical sensitivity analysis, identify the approaches for handling missing data that were considered and pros and cons of each*)

For these variables, rates of missing data are typically less than 1% of the state database. It is unlikely the bias would occur from such a low rate of missing data.

2b7.3. What is your interpretation of the results in terms of demonstrating that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias**?** (i.*e., what do the results mean in terms of supporting the selected approach for missing data and what are the norms for the test conducted; if no empirical analysis, provide rationale for the selected approach for missing data*)

Exclusion of cases for missing data is appropriate.

3. Feasibility

Extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

3a. Byproduct of Care Processes

For clinical measures, the required data elements are routinely generated and used during care delivery (e.g., blood pressure, lab test, diagnosis, medication order).

3a.1. **Data Elements Generated as Byproduct of Care Processes.**

Coded by someone other than person obtaining original information (e.g., DRG, ICD-9 codes on claims)

If other:

3b. Electronic Sources

The required data elements are available in electronic health records or other electronic sources. If the required data are not in electronic health records or existing electronic sources, a credible, near-term path to electronic collection is specified.

3b.1. **To what extent are the specified data elements available electronically in defined fields?** *(i.e., data elements that are needed to compute the performance measure score are in defined, computer-readable fields)*

ALL data elements are in defined fields in electronic claims

3b.2. **If ALL the data elements needed to compute the performance measure score are not from electronic sources, specify a credible, near-term path to electronic capture, OR provide a rationale for using other than electronic sources.**

3b.3. **If this is an eMeasure, provide a summary of the feasibility assessment in an attached file or make available at a measure-specific URL.**

Attachment:

3c. Data Collection Strategy

Demonstration that the data collection strategy (e.g., source, timing, frequency, sampling, patient confidentiality, costs associated with fees/licensing of proprietary measures) can be implemented (e.g., already in operational use, or testing demonstrates that it is ready to put into operational use). For eMeasures, a feasibility assessment addresses the data elements and measure logic and demonstrates the eMeasure can be implemented or feasibility concerns can be adequately addressed.

3c.1. **Describe what you have learned/modified as a result of testing and/or operational use of the measure regarding data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, time and cost of data collection, other feasibility/implementation issues.**

IF a PRO-PM, consider implications for both individuals providing PROM data (patients, service recipients, respondents) and those whose performance is being measured.

Because the indicator is based on readily available administrative billing and claims data and U.S. Census data, feasibility is not an issue.

The AHRQ QI software has been publicly available at no cost since 2001; Users have over ten years of experience using the AHRQ QI software in SAS and Windows.

3c.2. **Describe any fees, licensing, or other requirements to use any aspect of the measure as specified** *(e.g., value/code set, risk model, programming code, algorithm)*.

There are no fees. Software is freely available from the AHRQ Quality Indicators website (http://www.qualityindicators.ahrq.gov/).

4. Usability and Use

Extent to which potential audiences (e.g., consumers, purchasers, providers, policy makers) are using or could use performance results for both accountability and performance improvement to achieve the goal of high-quality, efficient healthcare for individuals or populations.

4a. Accountability and Transparency

Performance results are used in at least one accountability application within three years after initial endorsement and are publicly reported within six years after initial endorsement (or the data on performance results are available). If not in use at the time of initial endorsement, then a credible plan for implementation within the specified timeframes is provided.

4.1. **Current and Planned Use**

NQF-endorsed measures are expected to be used in at least one accountability application within 3 years and publicly reported within 6 years of initial endorsement in addition to performance improvement.

Planned	Current Use (for current use provide URL)
Quality Improvement (Internal to the specific organization)	Public Reporting Wisconsin Hospital Association (WHA) Information Center, Wisconsin Inpatient Hospital Quality Indicators Report http://www.whainfocenter.com/uploads/PDFs/Publications/QualityIndicators/2012_WI_IQIReport.pdf Quality Improvement with Benchmarking (external benchmarking to multiple organizations) Norton Healthcare http://www.rivercityortho.com/ChildrenInfectionControl https://nortonhealthcare.com/Pages/Home.aspx

4a.1. **For each CURRENT use, checked above, provide:**

- Name of program and sponsor
- Purpose
- Geographic area and number and percentage of accountable entities and patients included

Public Reporting:
Wisconsin Hospital Association (WHA) Information Center
Wisconsin Hospital Association (WHA) Quality Indicators Report
http://www.whainfocenter.com/uploads/PDFs/Publications/QualityIndicators/2012_WI_IQIReport.pdf

Quality Improvement:

Norton Healthcare
Large healthcare provider in Kentucky and Southern Indiana that includes 5 hospitals and 12 immediate care centers.
http://www.rivercityortho.com/ChildrenInfectionControl
https://nortonhealthcare.com/Pages/Home.aspx

4a.2. **If not currently publicly reported OR used in at least one other accountability application (e.g., payment program, certification, licensing) what are the reasons?** (*e.g., Do policies or actions of the developer/steward or accountable entities restrict access to performance results or impede implementation?*)
n/a

4a.3. **If not currently publicly reported OR used in at least one other accountability application, provide a credible plan for implementation within the expected timeframes -- any accountability application within 3 years and publicly reported within 6 years of initial endorsement.** (*Credible plan includes the specific program, purpose, intended audience, and timeline for implementing the measure within the specified timeframes. A plan for accountability applications addresses mechanisms for data aggregation and reporting.*)
n/a

4b. Improvement

Progress toward achieving the goal of high-quality, efficient healthcare for individuals or populations is demonstrated. If not in use for performance improvement at the time of initial endorsement, then a credible rationale describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.

4b.1. **Progress on Improvement. (Not required for initial endorsement unless available.)**
Performance results on this measure (current and over time) should be provided in 1b.2 and 1b.4. Discuss:

- **Progress (trends in performance results, number and percentage of people receiving high-quality healthcare)**
- **Geographic area and number and percentage of accountable entities and patients included**

Overall and average hospital performance has improved from 24.0 in 2011 to 19.4 in 2013 nationwide. For hospitals, the average rate has also decreased from 11.5 in 2011 to 9.15 in 2013, and the variation in hospital rates has also decreased with a standard deviation of 22.6 vs. 19.1. However, because we can only examine 3 years with adequate present on admission data we cannot determine if the decrease observed in 2013 will be sustained in future years.

See Table 1 (section 1b.2., also included in supplemental materials)

4b.2. **If no improvement was demonstrated, what are the reasons? If not in use for performance improvement at the time of initial endorsement, provide a credible rationale that describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.**
n/a

4c. Unintended Consequences

The benefits of the performance measure in facilitating progress toward achieving high-quality, efficient healthcare for individuals or populations outweigh evidence of unintended negative consequences to individuals or populations (if such evidence exists).

4c.1. **Were any unintended negative consequences to individuals or populations identified during testing; OR has evidence of unintended negative consequences to individuals or populations been reported since implementation? If so, identify the negative unintended consequences and describe how benefits outweigh them or actions taken to mitigate them.**
No evidence has been identified suggesting unintended consequences for this measure.

5. Comparison to Related or Competing Measures

If a measure meets the above criteria and there are endorsed or new related measures (either the same measure focus or the same target population) or competing measures (both the same measure focus and the same target population), the measures are compared to address harmonization and/or selection of the best measure.

5. Relation to Other NQF-endorsed Measures
Are there related measures (conceptually, either same measure focus or target population) or competing measures (conceptually both the same measure focus and same target population)? If yes, list the NQF # and title of all related and/or competing measures.
Yes

5.1a. **List of related or competing measures (selected from NQF-endorsed measures)**
1731 : PC-04 Health Care-Associated Bloodstream Infections in Newborns

5.1b. **If related or competing measures are not NQF endorsed please indicate measure title and steward.**

5a. Harmonization

The measure specifications are harmonized with related measures;
OR
The differences in specifications are justified

5a.1. **If this measure conceptually addresses EITHER the same measure focus OR the same target population as NQF-endorsed measure(s):**
Are the measure specifications completely harmonized?

5a.2. **If the measure specifications are not completely harmonized, identify the differences, rationale, and impact on interpretability and data collection burden.**

5b. Competing Measures

The measure is superior to competing measures (e.g., is a more valid or efficient way to measure);
OR
Multiple measures are justified.

5b.1. **If this measure conceptually addresses both the same measure focus and the same target population as NQF-endorsed measure(s):**
Describe why this measure is superior to competing measures (e.g., a more valid or efficient way to measure quality); OR provide a rationale for the additive value of endorsing an additional measure. (Provide analyses when possible.)
Our understanding is that The Joint Commission (TJC) intents to submit "Health Care-Associated Bloodstream Infections in Newborns (PC-04)" under the call for measures. In anticipation of this, AHRQ and TJC have agreed to harmonize our measures to the extent feasible given alternative data sources. (The AHRQ QI is an existing NQF endorsed measure; the TJC measure is a newly submitted measure).

There are three specification differences related to data availability in the TJC measure specification. First, hospitals report to TJC the actual birth weight from the medical record (rather than coded birth weight using ICD-9-CM); Second, hospitals report whether the patient has a signed consent form for participation in a clinical trial. Therefore, the TJC specification does not include an inclusion criteria related to gestational age as in the AHRQ QI (rather, actual birthweight is used as an alternative to coded birth weight). The TJC also includes an exclusion for enrollment in a clinical trial. The AHRQ QI contains no such exclusion. Finally, TJC excludes stays of more than 120 days for technical reasons related to the measure reporting period. This rationale does not apply to the AHRQ QI, and therefore the AHRQ QI has no such exclusion.

Appendix

A.1 Supplemental materials may be provided in an appendix. All supplemental materials (such as data collection instrument or

methodology reports) should be organized in one file with a table of contents or bookmarks. If material pertains to a specific submission form number, that should be indicated. Requested information should be provided in the submission form and required attachments. There is no guarantee that supplemental materials will be reviewed.

Attachment **Attachment:** NQI03_Supplemental_Files_160216.pdf

Contact Information

Co.1 **Measure Steward (Intellectual Property Owner):** Agency for Healthcare Research and Quality
Co.2 **Point of Contact:** Mia, DeSoto, Maushami.Desoto@ahrq.hhs.gov
Co.3 **Measure Developer if different from Measure Steward:**
Co.4 **Point of Contact:**

Additional Information

Ad.1 **Workgroup/Expert Panel involved in measure development**
Provide a list of sponsoring organizations and workgroup/panel members' names and organizations. Describe the members' role in measure development.
Members of this workgroup have provided feedback on key indicator development decisions and methodology, including a review of validity and usefulness of Neonatal Blood Stream Infection Rate (NQI 03).

J. Christopher Glantz, MD, MPH, OB/GYN, Maternal-Fetal Medicine
Rochester, New York
Strong Memorial Hospital (University of Rochester Medical Center)
Nominated by the American Academy of Pediatrics

William G. Keyes, MD, PhD, Pediatrics, Pediatric Critical Care Medicine, Neonatal-Perinatal Medicine
Atlanta, Georgia
Children's Healthcare of Atlanta, Northside Hospital
Nominated by the National Association of Children´s Hospitals and Related Institutions

Teresa W Marchese, CNM, PhD, Nurse-Midwife
Washington, DC
Unity Health Care, Inc
Georgetown University School of Nursing and Health Studies
Nominated by the American College of Nurse-Midwives

Richard A. Molteni, MD, FAAP, Neonatal Medicine, Pediatrics
Seattle, Washington
Children's Hospital and Regional Medical Center, Seattle
Nominated by the Child Health Corporation of America

Paul Ogburn, Jr, MD, OB/GYN, Maternal-Fetal Medicine
East Setauket, New York
SUNY University Hospital Stony Brook
Nominated by the American Academy of Pediatrics

Sumana Reddy, MD, FAAFP, Family Medicine, Obstetrics
Salinas, CA
Salinas Valley Memorial and Natividad Medical Centers
Nominated by the California Academy of Family Physicians

William F. Walsh, MD, Pediatrics, Neonatology
Nashville, Tennessee
Vanderbilt Children's Hospital
Nominated by the National Association of Children´s Hospitals and Related Institutions

Lorna Cisler-Cahill, MS, RN, Neonatal Clinical Nurse Specialist
Milwaukee, Wisconsin
Children's Hospital of Wisconsin
Nominated by the Child Health Corporation of America

Measure Developer/Steward Updates and Ongoing Maintenance

Ad.2 **Year the measure was first released:** 2008

Ad.3 **Month and Year of most recent revision:** 01, 2015

Ad.4 **What is your frequency for review/update of this measure?** Annual

Ad.5 **When is the next scheduled review/update for this measure?** 01, 2016

Ad.6 **Copyright statement:** The AHRQ QI software is publicly available. We have no copyright disclaimers.

Ad.7 **Disclaimers:** None

Ad.8 **Additional Information/Comments:** None

MEASURE WORKSHEET

This document summarizes the evaluation of the measure as it progresses through NQF's Consensus Development Process (CDP). The information submitted by measure developers/stewards is included after the Brief Measure Information, Preliminary Analysis, and Pre-meeting Public and Member Comments sections.

To navigate the links in the worksheet: Ctrl + click link to go to the link; ALT + LEFT ARROW to return

Brief Measure Information

NQF #: 0480
Measure Title: PC-05 Exclusive Breast Milk Feeding
Measure Steward: The Joint Commission
Brief Description of Measure: PC-05 assesses the number of newborns exclusively fed breast milk during the newborn´s entire hospitalization. This measure is a part of a set of five nationally implemented measures that address perinatal care (PC-01: Elective Delivery, PC-02: Cesarean Birth, PC-03: Antenatal Steroids, PC-04: Health Care-Associated Bloodstream Infections in Newborns).
Developer Rationale: Increasing the number of newborns who are exclusively fed breast milk for the first six months of life remains a major goal of the WHO, DHHS, AAP and ACOG. Guidelines for the promotion of breast milk feeding are available from the CDC to assist hospitals in establishing successful interventions to improve exclusive breast milk feeding rates in newborns. Breast milk feeding results in numerous health benefits for both mother and newborn. Breastfeeding is associated with decreased risk for many early-life diseases and conditions, including otitis media, respiratory tract infections, atopic dermatitis, gastroenteritis, type 2 diabetes, sudden infant death syndrome, and obesity. Breastfeeding also is associated with health benefits to women, including decreased risk for type 2 diabetes, ovarian cancer, and breast cancer
The measure will assist health care organizations (HCOs) to track evidence of an increase in the number of newborns who were exclusively fed breast milk during the birth hospitalization.

Numerator Statement: Newborns that were fed breast milk only since birth
Denominator Statement: Single term liveborn newborns discharged alive from the hospital with ICD-10-CM Principal Diagnosis Code for single liveborn newborn as defined in Appendix A, Table 11.20.1 available at: http://manual.jointcommission.org/releases/TJC2015B2/
Denominator Exclusions: • Admitted to the Neonatal Intensive Care Unit (NICU) at this hospital during the hospitalization
• ICD-10-CM Other Diagnosis Codes for galactosemia as defined in Appendix A, Table 11.21
• ICD-10-PCS Principal Procedure Code or ICD-10-PCS Other Procedure Codes for parenteral infusion as defined in Appendix A, Table 11.22
• Experienced death
• Length of Stay >120 days
• Enrolled in clinical trials
• Patients transferred to another hospital
• Patients who are not term or with < 37 weeks gestation completed

Measure Type: Process
Data Source: Electronic Clinical Data, Paper Medical Records
Level of Analysis: Facility, Population : National

IF Endorsement Maintenance – Original Endorsement Date: Oct 24, 2008 **Most Recent Endorsement Date:** Mar 30, 2012

Maintenance of Endorsement -- Preliminary Analysis

To maintain NQF endorsement endorsed measures are evaluated periodically to ensure that the measures still meets the NQF endorsement criteria ("maintenance"). The emphasis for maintaining endorsement is focused on how effective the measure is for promoting improvements in quality. Endorsed measures should have some experience from the field to inform the evaluation. The emphasis for maintaining endorsement is noted for each criterion.

Criteria 1: Importance to Measure and Report

1a. Evidence
Maintenance measures – less emphasis on evidence unless there is new information or change in evidence since the prior evaluation.

1a. Evidence. The evidence requirements for a *process or intermediate outcome* measure is that it is based on a systematic review (SR) and grading of the body of empirical evidence where the specific focus of the evidence matches what is being measured.

The developer provides the following evidence for this measure:

- **Systematic Review of the evidence specific to this measure?** ☒ **Yes** ☐ **No**
- **Quality, Quantity and Consistency of evidence provided?** ☒ **Yes** ☐ **No**
- **Evidence graded?** ☐ **Yes** ☒ **No**

Summary of prior review in 2012

The evidence supporting this measure is based on clinical practice guidelines from the Academy of Breastfeeding Medicine, based on recommendations from the Office on Women´s Health of the U.S. Department of Health and Human Services, the American Academy of Pediatrics, the American College of Obstetricians and Gynecologists, the American Academy of Family Physicians, the World Health Organization (WHO), the Academy of Breastfeeding Medicine, and the UNICEF/WHO evidence-based "Ten Steps to Successful Breastfeeding." The recommendations were based on a review of the literature, which indicates that there are numerous health benefits to breastfeeding for both mother and newborn. The developer states that over 27,000 articles on this topic were published between 1980 and 2012, and the literature includes "900 studies which examine outcomes from breast-feeding with reductions in asthma, diarrheal illness, and childhood obesity being the most important health benefits."

Changes to evidence from last review

☒ **The developer attests that there have been no changes in the evidence since the measure was last evaluated.**
☐ **The developer provided updated evidence for this measure:**

Updates:

Guidance from the Evidence Algorithm

Process measure (Box 1) → Systematic review (Box 3) → QQC present (Box 4) →SR concludes high quality evidence

Questions for the Committee:

- *Does the Committee agree the evidence basis for the measure has not changed and there is no need for repeat discussion and voting on Evidence?*

Preliminary rating for evidence: ☒ **Pass** ☐ **No Pass**

1b. Gap in Care/Opportunity for Improvement and 1b. Disparities
Maintenance measures – increased emphasis on gap and variation

1b. Performance Gap. The performance gap requirements include demonstrating quality problems and opportunity for

improvement.

The developer believes a goal of 70% for exclusive breastmilk feeding is achievable. Rates for exclusive breast milk feeding remain below 50% for over half of the hospitals reporting data. Only the 90th percentile hospitals (74.3%) are performing above the goal of 70%. The developer reports the following data:

Baseline 2Q 2010: 165 hospitals, average measure rate of 40.9% (n=54,630 patients).

Beginning 2014, all hospitals with >1110 births/year were required to report; 1386 hospitals reported the data with an average rate of 49.4% (n=728,157 patients).

	2011	2012	2013	2014
# hospitals	166	170	197	1386
# newborns	69,613	76,952	91,011	728,157
National aggregate rate	45.9%	51.1%	53.6%	49.4%
10th-25th -90th %tile	19.9% - 34.8% - 62.8%	21.1% - 41.7% - 80.7%	28.4% - 42.3% - 79.3%	22.0% - 35.1% - 74.3%
Mean hospital rate (SD)	0.49335 (0.21348)	0.55872 (0.20359)	0.5632 (0.19707)	0.48724 (0.19475)

Beginning in January 2016 hospitals with >300 births/year are required to report – an additional 821 (approximately 80% of all birthing hospitals).

Disparities

No disparities information from the use of the measure is provided. The developer provides literature which states that in a study of 307 mothers, "exclusive in-hospital breast milk feeding was reported by 54.2% of white, 38.7% of black, 54.0% of Asian, and 44.7% of Hispanic ($p = 0.063$), and among these, only 55.6%, 50.0%, 58.9%, and 19.1%, respectively, maintained exclusive breast milk feeding during the first postpartum month ($p < 0.02$). The rate of exclusive breast milk feeding at the end of the first month was 10.5%, 15.8%, 20.7%, and 3.9%, respectively, for the white, black, Asian, and Hispanic mothers whose infants received partial or no breastfeeding in-hospital."

The developer also provided CDC data stating that "from 2000-2004 the rates of exclusive breastfeeding were significantly lower among black infants (compared with white infants) and infants born to unmarried mothers (compared with married mothers). Additionally, older age, urban residence, higher education, and higher income of mothers all were positively associated with exclusive breast milk feeding."

Questions for the Committee:

- *Is there a gap in care that warrants a national performance measure?*
- *Can this measure be used to understand disparities in perinatal care?*

Preliminary rating for opportunity for improvement: ☒ **High** ☐ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 1: Importance to Measure and Report (including 1a, 1b, 1c)

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability

2a1. Reliability Specifications
Maintenance measures – no change in emphasis – specifications should be evaluated the same as with new measures

2a1. Specifications requires the measure, as specified, to produce consistent (reliable) and credible (valid) results about the quality of care when implemented.

Data source(s): Electronic clinical data, paper medical records

Specifications:

- The measure has been converted to ICD-10 codes.
- The sub-measure PC-05a: Exclusive Breast Milk Feeding Considering Mother's Initial Feeding Plan was retired. The developer states this was "based on feedback from key stakeholders – including health care organizations, The Centers for Disease Control and Prevention (CDC), the American College of Obstetricians and Gynecologists (ACOG), the American Academy of Pediatrics (AAP) and the Association of Women's Health, Obstetric & Neonatal Nurses (AWHONN). The sub-measure greatly increased the burden of data abstraction, did not follow normal workflow patterns and impeded improved performance for PC-05."

  Staff note: In 2012 an appeal was lodged against endorsement of this measure by the International Formula Council citing lack of exclusions for women who choose not to breastfeed. In response, the developer offered to include an additional stratified rate that excluded mothers that choose not to breastfeed.

- The denominator data element Reason for Not Exclusively Feeding Breast Milk was removed as a denominator exclusion, since this concept cannot be modeled in the eCQM version of the measure and removal greatly reduces the burden of data abstraction.
- Late preterm newborns are now excluded from the denominator population with the new data element Term Newborn instead of using diagnosis codes for premature newborns found on Table 11.23 in Appendix A. This change was made because prematurity codes were not being routinely assigned unless birth weight codes are also assigned.
- There are 11 data elements used to calculate the denominator: Admission date; admission to the NICU; birthdate; clinical trial; discharge date; discharge disposition; ICD-10-CM Other Diagnosis Codes; ICD-10-PCS Other Procedure Codes; ICD-10-CM Principal Diagnosis Code; ICD-10-CM Principal Procedure Code; term newborn.
- Sampling is allowed.
- A calculation algorithm is included.

Questions for the Committee:

- *Is the removal of sub-measure PC-05a reasonable?*
- *Are the changes to the denominator appropriate?*
- *Are all the data elements clearly defined? Are all appropriate codes included?*
- *Is the logic or calculation algorithm clear?*
- *Is it likely this measure can be consistently implemented?*

2a2. Reliability Testing Testing attachment
Maintenance measures – less emphasis if no new testing data provided

2a2. Reliability testing demonstrates if the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise enough to distinguish differences in performance across providers.

For maintenance measures, summarize the reliability testing from the prior review:

Inter-rater reliability (IRR) was performed by ORXY vendor reabstaraction for 106 hospitals comprising 26,302 records. IRR is an appropriate method of assessing data element reliability for chart abstraction. The agreement rate for the data element "Exclusive breastmilk feeding" was 97.53%.

Describe any updates to testing
None

SUMMARY OF TESTING
Reliability testing level ☐ **Measure score** ☒ **Data element** ☐ **Both**
Reliability testing performed with the data source and level of analysis indicated for this measure ☒ **Yes** ☐ **No**

Method(s) of reliability testing see above

Results of reliability testing see above Only % agreement, no statistical tests such as Kappa or ICC was provided.

Guidance from the Reliability Algorithm

Guidance from the Reliability Algorithm
Precise specifications (Box 1) → Empirical reliability testing (Box 2) → Data element testing (Box 8) → appropriate method- IRR (Box 9) → high or moderate confidence of reliability of numerator data element → moderate (highest rating possible)

Questions for the Committee:
- *The developer does not provide any additional testing for reliability since the last NQF endorsement review. Does the Committee agree there is no need for repeat discussion and voting on Evidence?*

Preliminary rating for reliability: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

2b. Validity
Maintenance measures – less emphasis if no new testing data provided

2b1. Validity: Specifications

2b1. Validity Specifications. This section should determine if the measure specifications are consistent with the evidence.
Specifications consistent with evidence in 1a. ☒ **Yes** ☐ **Somewhat** ☐ **No**

Question for the Committee:
- *Are the specifications consistent with the evidence?*

2b2. Validity testing

2b2. Validity Testing should demonstrate the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality.

For maintenance measures, summarize the validity testing from the prior review:
Prior testing was face validity only

Describe any updates to validity testing
New empirical validity testing of the measure score is provided.

SUMMARY OF TESTING
Validity testing level ☒ **Measure score** ☐ **Data element testing against a gold standard** ☐ **Both**

Method of validity testing of the measure score:
- ☐ **Face validity only**
- ☒ **Empirical validity testing of the measure score**

Validity testing method:

- 1,352 hospitals submitted 775,909 inpatient records from one year (Q3 2014- Q2 2015). Measure convergent validity testing was performed using patient level data. The Spearman rank-order correlation (a nonparametric measure of association based on the ranks of the data values by measure PC-05 and hospitals) was used to correlate the results from this measure with other measures in the Joint Commission's perinatal set. The developer hypothesizes that hospitals that perform well on this measure will perform well on all measures in the set.

Validity testing results:

- The developer provides scatter plots and a correlation table. They summarize the results: "the correlation of PC-05 with PC-02 is moderate and statistically significant. The other correlations with the other PC measures are relatively weak and not significant. Performance of hospitals on this measure varied widely, 90% of the hospital measure rates fall between 23 and 75%, indicating that there is much room for improvement on this measure."

Questions for the Committee:

- *Does the Committee expect strong correlations in performance among the five perinatal measures?*
- *Is the test sample adequate to generalize for widespread implementation?*
- *Do the results demonstrate sufficient validity so that conclusions about quality can be made?*
- *Do you agree that the score from this measure as specified is an indicator of quality?*

2b3-2b7. Threats to Validity

2b3. Exclusions:

Frequency of exclusions were presented:

1,352 hospitals submitted 775,909 inpatient records

Data element	Overall %	Median%	Min - Max	
Admitted to the NICU	5.25%	4.25%	0-69%	
ICD-9-CM Other Diagnosis Codes for galactosemia	0 occurrences	0	0	
ICD-9-CM Principal Procedure Code or ICD-9-CM Other Procedure Codes for parenteral infusion	0 occurrences	0	0	
Patients who expire during the hospital stay	0.05%	0%	0-1.9%	
Length of Stay >120 days	0 occurrences	0	0	
Clinical Trials	0.03%	0%	0-32%	
Documented Reason for Not Exclusively Feeding Breast Milk	0.94%	0.6%	0-17.2%	
Patients transferred to another hospital;	0.06%	0%	0-5.5%	
ICD-9-CM Other Diagnosis Codes for premature newborns	0 occurrences	0	0	

Questions for the Committee:

- *Are the exclusions consistent with the evidence?*
- *Are any patients or patient groups inappropriately excluded from the measure?*
- *Are the exclusions/exceptions of sufficient frequency and variation across providers to be needed (and outweigh the data collection burden)?*

2b4. Risk adjustment: **Risk-adjustment method** ☒ **None** ☐ **Statistical model** ☐ **Stratification**

2b5. Meaningful difference *(can statistically significant and clinically/practically meaningful differences in performance measure scores can be identified)*:

The developer reports the following descriptive analysis of the measure results:
N = 1,352 hospitals Number of records =775,909
Percentile 10%: 23%
Percentile 25%: 37%
Median: 50%
Percentile 75%: 63%
Percentile 90%: 75%

The Joint Commission's Target Analysis uses two methods: Given the national average for a performance measure, a target range is constructed. Using generalized linear mixed models methodology (also known as hierarchical models), a predicted estimate of an HCO's performance, with a corresponding 95% confidence interval, is generated. This confidence interval is compared to the target range, to determine the HCOs' rating. The estimate of the organization's true performance is based on both the data from that organization and on data from the entire set of reporting organizations. The object of target analysis is to compare a health care organizations (HCO) data against a comparative norm for the purpose of evaluating performance improvement opportunities. When an organization's performance level is statistically significantly different from a comparative norm, it is considered a statistical deviation. A statistical deviation may be desirable or undesirable depending on the "direction of improvement" of the measure.

PC-05 Distribution of Outliers 2011 1st Quarter Data:
Scores on this measure: N=161, Mean 48.33%, SD 0.23493
10th Percentile= 19.23%
25th Percentile= 31.88%
50th Percentile= 50%
75th Percentile= 63.6%
90th Percentile= 78.95%

4 (2.48%) Favorable – results statistically significantly higher than the national rate
119 (73.91%) Neutral – results not significantly different from target range
38 (23.6%) Undesirable –results statistically significantly lower than the national rate

Question for the Committee:

- *Does this measure identify meaningful differences about quality?*

2b6. Comparability of data sources/methods:

N/A

2b7. Missing Data

The developer reports "Any file with missing data will result in a measure category assignment of X and rejection of the file from the warehouse, unless the data are not used to process the measure. If the data are used to process the measure and have been reported by the abstractor as No/UTD, the case will result in a measure category assignment of E (i.e., failed measure, not rejected."

Guidance from algorithm: Specifications consistent with evidence (Box 1) → Assessed all potential threats to validity

(Box 2) → empirical testing (Box 3) →testing of measure score (Box 6) →appropriate testing method (Box 7) → testing results moderate (or possible low) → moderate

Preliminary rating for validity: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 2: Scientific Acceptability of Measure Properties (including all 2a, 2b, and 2d)

Criterion 3. Feasibility

Maintenance measures – no change in emphasis – implementation issues may be more prominent

3. Feasibility is the extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

- According to the developer, "hospitals using this performance measure generally collect measure data via manual review of the paper medical record, the EMR or a combination of both."
- They also report "not all hospitals currently have the capacity to abstract the electronic version of this measure, so continues to offer this chart abstracted version which allows for data capture from unstructured data fields. All data elements needed to compute the PC-05 performance measure score have been retooled for capture from electronic sources. Annual updates are performed to match the eCQM specifications to the current version of the chart-abstracted specifications."
- The measure specifications are freely available and there are no fees or licensing requirements.

Questions for the Committee:

- *What is the burden for hospitals in collecting data for this measure?*
- *Are the required data elements routinely generated and used during care delivery?*
- *What is the burden of data collection for this measure?*
- *Are the required data elements available in electronic form, e.g., EHR or other electronic sources?*

Preliminary rating for feasibility: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 3: Feasibility

Criterion 4: Usability and Use

Maintenance measures – increased emphasis – much greater focus on measure use and usefulness, including both impact /improvement and unintended consequences

4. Usability and Use evaluate the extent to which audiences (e.g., consumers, purchasers, providers, policymakers) use or could use performance results for both accountability and performance improvement activities.

Current uses of the measure [from OPUS]

Publicly reported? ☒ **Yes** ☐ **No**

Current use in an accountability program? ☒ **Yes** ☐ **No**

OR

Planned use in an accountability program? ☐ **Yes** ☐ **No**

Accountability program details

This measure is currently in use in the following programs:

Quality Check – Publically available website with performance data on 3,300 Joint Commission-accredited hospitals

http://www.qualitycheck.org/consumer/searchQCR.aspx

Hospital Accreditation Program – The Joint Commission's accreditation program (3,300 hospitals, nationwide) http://jointcommission.org

The Joint Commission's Annual Report, America's Hospitals: Improving Quality and Safety (3,300 hospitals, nationwide) http://www.jointcommission.org/annualreport.aspx

Perinatal Care Certification – The Joint Commission's certification for "hospitals that have achieved integrated, coordinated, patient-centered care for clinically uncomplicated pregnancies and births" (12 hospitals nationwide in 2016) http://www.jointcommission.org/certification/perinatal_care_certification.aspx

Improvement results

The developer reports: "The rate of exclusive breast milk feeding slowly improved from 40.9% in 2010 with 165 hospitals reporting to 49.4% in 2014 with 1386 hospitals reporting based on Joint Commission ORYX performance measurement data. Beginning with January 1, 2016 discharges, an additional 821 accredited hospitals will begin reporting the data. The new reporting requirement will capture approximately 80% of the accredited hospitals with maternity services in the US. The most improvement in 2014 was noted for the 90th percentile (74.3%) hospitals. This underscores the importance to continue to monitor progress towards improving the rate in order to reach the performance goal of 70%."

Unexpected findings (positive or negative) during implementation

The developer notes three unintended consequences:

- Data abstraction for the mother's initial feeding plan for PC-05a: Exclusive Breast Milk Feeding Considering Mother's Initial Feeding Plan did not follow normal workflow patterns and greatly increased the burden of data abstraction for hospitals. Numerous stakeholders also expressed concern that undecided mothers were often choosing both formula and breast milk, and it was perceived that these mothers were not receiving the same level of support as mothers who had chosen to feed breast milk only. ACTION: Retired the sub-measure PC-05a: Exclusive Breast Milk Feeding Considering Mother's Initial Feeding Plan.
- Late preterm newborns were not being routinely excluded from the denominator population via diagnosis codes for premature newborns on Table 11.23. ACTION: Table 11.23 was removed and a new denominator data element Term Newborn was added, so that now only newborns who were term or =>37 weeks gestation completed were included.
- Reviewing medical records for maternal medical conditions as reasons for not exclusively feeding breast milk was greatly increasing the burden of data abstraction based on feedback from hospitals. ACTION: The denominator data element Reason For Not Exclusively Feeding Breast Milk was removed from the denominator excluded population in order to simplify data abstraction.

Potential harms

Feedback: In 2013 MAP supported this measure for the Hospital Inpatient Quality Reporting Program.

***Questions for the Committee**:*

- *Committee members who are using /have used this measure should share their experiences. Are you aware of any unintended consequences of this measure?*
- *Are the revisions to the measure resulting from the mitigating actions appropriate?*
- *How can the performance results be used to further the goal of high-quality, efficient healthcare?*
- *Do the benefits of the measure outweigh any potential unintended consequences?*

Preliminary rating for usability and use: ☒ **High** ☐ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 4: Usability and Use

Criterion 5: Related and Competing Measures

Related or competing measures

- There are no related or competing measures.

Harmonization

N/A

Pre-meeting public and member comments

- I am writing to express my strong support for continued endorsement of PC-05, Exclusive Breast Milk Feeding. Seminal research from the California Maternity Care Quality Collaborative demonstrated substantial variation in supplementation of breastfed infants among maternity centers. Moreover, national data confirm that there is wide variation in the use of formula among breastfed infants in the first 2 days of life[1], ranging from 6.1% in Montana to 34.9% in New Jersey. This variation among states suggests that overutilization of formula occurs in many maternity hospitals.
  However, just as some infants require delivery via cesarean, some breastfeeding dyads require formula supplementation. It is therefore essential that implementation of PC05 occurs within a context that provides appropriate support for family-centered decision-making and transitions to outpatient support. The AAP recommends that all breastfeeding newborns be seen within 48 to 72h of discharge from the maternity center[2]. Because some families may initiate breastfeeding after leaving the hospital, it may be prudent to schedule all newborns for a 48 to 72h visit to establish care with a pediatric provider. It may be useful to consider a quality measure for the proportion of infants seen by a health professional, either in the office or for a home visit, within 48 to 72h of discharge.
  Of note, the Baby Friendly Hospital Initiative includes a metric for exclusive breast milk feeding as one of its metrics for certification. Differences exist between PC-05 and the BFHI measure, increasing reporting burden for maternity centers. It would be helpful if BFHI and NQF could work together to develop a common metric for measuring exclusive breast milk feeding.
  Evidence continues to accrue that there is no replacement for mother's milk[3]. We can enable families to achieve optimal infant feeding by reducing iatrogenic formula supplementation during the maternity stay, and by ensuring careful follow-up for all families in the early days of life.
  1. Centers for Disease Control and Prevention. 2012: Percent of breastfed infants who were supplemented with infant formula within 2 days of life. 2015 [cited 2016 April 4]; Available from: https://nccd.cdc.gov/NPAO_DTM/IndicatorSummary.aspx?category=8&indicator=41.
  2. American Academy of Pediatrics, Breastfeeding and the use of human milk. Pediatrics, 2012. 129(3): p. e827-41.
  3. Victora, C.G., et al., Breastfeeding in the 21st century: epidemiology, mechanisms, and lifelong effect. The Lancet, 2016. 387(10017): p. 475-490.
- **Developer Response:**
  The Joint Commission thanks you for your support of NQF measure # 0480 PC-05: Exclusive Breast Milk Feeding. We have reached out to Baby Friendly USA in order to get a better understanding of the differences in their exclusive breast milk feeding measure specifications and our measure specifications with a goal of harmonizing to the extent possible.

NATIONAL QUALITY FORUM

Measure missing data in MSF 6.5 from MSF 5.0

NQF #: 0480 **NQF Project:** Perinatal and Reproductive Health Project

1. IMPACT, OPPORTUITY, EVIDENCE - IMPORTANCE TO MEASURE AND REPORT

Importance to Measure and Report is a threshold criterion that must be met in order to recommend a measure for endorsement. All three subcriteria must be met to pass this criterion. See guidance on evidence.
Measures must be judged to be important to measure and report in order to be evaluated against the remaining criteria.
(evaluation criteria)

1c.1 **Structure-Process-Outcome Relationship** *(Briefly state the measure focus, e.g., health outcome, intermediate clinical outcome, process, structure; then identify the appropriate links, e.g., structure-process-health outcome; process- health outcome; intermediate clinical outcome-health outcome)*:
The focus of the measure is to increase the number of newborns who are exclusively fed breast milk during the birth hospitalization >> population determined >> population assessed >> newborns exclusively fed breast milk while in the hospital >> reduced morbidity and mortality of for mother and newborn.

1c.2-3 **Type of Evidence** *(Check all that apply)*:
Clinical Practice Guideline
Systematic review of body of evidence (other than within guideline development)

1c.4 **Directness of Evidence to the Specified Measure** *(State the central topic, population, and outcomes addressed in the body of evidence and identify any differences from the measure focus and measure target population)*:
The central topic for the measure is promotion of exclusive breast milk feeding of the newborn during the entire birth hospitalization. The evidence shows numerous health benefits for both mothers and newborns. The target population for the performance measure is consistent with the body of evidence supporting the need for improving exclusive breast milk feeding rates.

1c.5 **Quantity of Studies in the Body of Evidence** *(Total number of studies, not articles)*: 1c.5. Quantity of Studies in the Body of Evidence (Total number of studies, not articles)
The body of literature examining breast feeding with neonatal outcomes is very large with over 27,000 articles published since 1980. 900 studies examine outcomes from breast-feeding with reductions in asthma, diarrheal illness, and childhood obesity being the most important health benefits. Exclusive breast-feeding in the first weeks was the single most important factor. Over 100 studies have examined initial breast feeding as a quality measure. A separate but related evidence base is the World Health Organization and United Nations Children´s Fund (UNICEF) Baby-Friendly Hospital Initiative that specifies Ten Steps to Successful Breastfeeding which identifies hospital practices that impair exclusive breast-feeding (over 200 separate studies).

1c.6 **Quality of Body of Evidence** (*Summarize the certainty or confidence in the estimates of benefits and harms to patients across studies in the body of evidence resulting from study factors. Please address: a) study design/flaws; b) directness/indirectness of the evidence to this measure (e.g., interventions, comparisons, outcomes assessed, population included in the evidence); and c) imprecision/wide confidence intervals due to few patients or events):* The quality of evidence supporting the promotion and support of exclusive breast milk feeding is quite high with studies published that have involved mother and newborn couplets. As noted numerous RCTs have been conducted over the past decades demonstrating improved health benefits for both mother and newborn. Some of the improved health benefits for newborns include: otitis media risk reduction by 23% (95% CI 9% to 36%), respiratory tract infections risk reduction by 72% (95% CI 46% to 86%), atopic dermatitis risk reduction by 42% (95% CI 8% to 59%), gastroenteritis risk reduction by 64% (95% CI 26% to 82%), type 2 diabetes risk reduction by 39 percent (95% CI 15% to 56%) , sudden infant death syndrome risk reduction by 36 percent (95% CI 19% to 49%), and obesity risk reduction in two studies by 7- 24% (95% CI 14% to 33% and 95% CI 1% to 12%)
No study design flaws were identified during the literature review.

1c.7 Consistency of Results <u>across Studies</u> *(Summarize the consistency of the magnitude and direction of the effect):* Studies spanning the past five decades have consistently demonstrated the health benefits of breast milk feeding for both mother and newborn. Again, some of the improved health benefits for newborns include: otitis media risk reduction by 23% (95% CI 9% to 36%), respiratory tract infections risk reduction by 72% (95% CI 46% to 86%), atopic dermatitis risk reduction by 42% (95% CI 8% to 59%), gastroenteritis risk reduction by 64% (95% CI 26% to 82%), type 2 diabetes risk reduction by 39 percent (95% CI 15% to 56%) , sudden infant death syndrome risk reduction by 36 percent (95% CI 19% to 49%), and obesity risk reduction in two studies by 7- 24% (95% CI 14% to 33% and 95% CI 1% to 12%)

1c.8 Net Benefit *(Provide estimates of effect for benefit/outcome; identify harms addressed and estimates of effect; and net benefit - benefit over harms)***:**
As described before, there are no known harms to patients associated with exclusive breast milk feeding. There are numerous studies documenting health benefits to both newborn and mother; therefore, the benefits of this recommended practice outweigh the harms.

1c.9 Grading of Strength/Quality of the Body of Evidence. Has the body of evidence been graded? No

1c.10 If body of evidence graded, identify the entity that graded the evidence including balance of representation and any disclosures regarding bias:

1c.11 System Used for Grading the Body of Evidence: Other

1c.12 If other, identify and describe the grading scale with definitions: Although grading of the evidence was not determined during our systematic review, it was determined that the guideline developers accounted for a balanced representation of information, looked beyond one specialty group or discipline, and provided information that was accessible and met the requirements set out in this measure maintenance form.

1c.13 Grade Assigned to the Body of Evidence: Not Applicable

1c.14 Summary of Controversy/Contradictory Evidence: There is no documented evidence regarding controversy about the benefits of exclusive breast milk feeding for mother and newborn.

1c.15 Citations for Evidence other than Guidelines*(Guidelines addressed below)*:

- American College of Obstetricians and Gynecologists (ACOG). (Feb. 2007). Committee on Obstetric Practice and Committee on Health Care for Underserved Women. Breastfeeding: Maternal and Infant Aspects. ACOG Committee Opinion 361.
- Centers for Disease Control and Prevention (CDC). (2011). Hospital support for breastfeeding: Preventing obesity begins in hospitals. CDC Vital Signs, Retrieved September 26, 2011 at: http://www.cdc.gov/VitalSigns/pdf/2011-08-vitalsigns.pdf
- Ip S, Chung M, Raman G, et al. Breastfeeding and maternal and infant health outcomes in developed countries. Rockville, MD: US Department of Health and Human Services; 2007. Retrieved on September 27, 2011 at: http://www.ahrq.gov/downloads/pub/evidence/pdf/brfout/brfout.pdf.
- Kramer, M.S. & Kakuma, R. (2002).Optimal duration of exclusive breastfeeding. [107 refs] Cochrane Database of Systematic Reviews. (1):CD003517.
- Shealy, K.R., Li, R., Benton-Davis, S., & Grummer-Strawn, L.M. (2005).The CDC guide to breastfeeding interventions. Atlanta, GA: US Department of Health and Human Services, CDC. Available at: http://www.cdc.gov/breastfeeding/pdf/breastfeeding_interventions.pdf
- US Department of Health and Human Services (DHHS). (2000). Healthy People 2010. Washington, DC. Retrieved on Setember 26, 2011 at: http://www.healthypeople.gov/2010
- US Department of Health and Human Services (DHHS). (2010). Healthy People 2020. Washington, DC. Retrieved on September 26, 2011 at: http://www.healthypeople.gov/2020
- World Health Organization (WHO). Indicators for assessing breastfeeding practices. Geneva, Switzerland: World Health Organization; 1991. Retrieved on September 27, 2011 at: http://www.who.int/child-adolescent-health/new_publications/nutrition/who_cdd_ser_91.14.pdf.

1c.16 Quote verbatim, <u>the specific guideline recommendation</u> *(Including guideline # and/or page #)***:**
The following major recommendations are included in the Academy of Breastfeeding Medicine Protocol # 7on pages 173-177:
Policy Statements
1. The "name of institution" staff will actively support breastfeeding as the preferred method of providing nutrition to infants. A multidisciplinary, culturally appropriate team comprising hospital administrators, physician and nursing staff, lactation consultants and

specialists, nutrition staff, other appropriate staff, and parents shall be established and maintained to identify and eliminate institutional barriers to breastfeeding. On a yearly basis, this group will compile and evaluate data relevant to breastfeeding support services and formulate a plan of action to implement needed changes. (III)

2. A written breastfeeding policy will be developed and communicated to all health care staff. The "name of institution" breastfeeding policy will be reviewed and updated biannually using current research as an evidence-based guide. (III)

3. All pregnant women and their support people as appropriate will be provided with information on breastfeeding and counseled on the benefits of breastfeeding, contraindications to breastfeeding, and risk of formula feeding (Academy of Breastfeeding Medicine Protocol Committee, "Clinical protocol #19," 2009). (II-1, II-2, III)

4. The woman´s desire to breastfeed will be documented in her medical record. (III)

5. Mothers will be encouraged to exclusively breastfeed unless medically contraindicated. The method of feeding will be documented in the medical record of every infant. (Exclusive breastfeeding is defined as providing breast milk as the sole source of nutrition.) Exclusively breastfed babies receive no other liquids or solids, with the exception of oral medications prescribed by a medical care provider for the infant.) (II-1, II-2, III)

6. At birth or soon thereafter all newborns, if baby and mother are stable, will be placed skin-to-skin with the mother. Skin-to-skin contact involves placing the naked baby prone on the mother´s bare chest. The infant and mother can then be dried and remain together in this position with warm blankets covering them as appropriate. Mother–infant couples will be given the opportunity to initiate breastfeeding within 1 hour of birth. Post-cesarean-birth babies will be encouraged to breastfeed as soon as possible, potentially in the operating room or recovery area (see Table 1 in the original guideline document). The administration of vitamin K and prophylactic antibiotics to prevent ophthalmia neonatorum should be delayed for the first hour after birth to allow uninterrupted mother–infant contact and breastfeeding (Academy of Breastfeeding Medicine Protocol Committee, "ABM clinical protocol #3," 2009; Mikiel-Kostyra, Mazur, & Boltruszko, 2002; Righard & Alade, 1990). (II-1)

7. Breastfeeding mother–infant couples will be encouraged to remain together throughout their hospital stay, including at night (rooming-in). Skin-to-skin contact will be encouraged as much as possible. (II-1)

8. Breastfeeding assessment, teaching, and documentation will be done on each shift and whenever possible with each staff contact with the mother. Each feeding will be documented, including latch, position, and any problems encountered, in the infant´s medical record. For feedings not directly observed, maternal report may be used. Every shift, a direct observation of the baby´s position and latch-on during feeding will be performed and documented. (II-1, II-2, III)

9. Mothers will be encouraged to utilize available breastfeeding resources including classes, written materials, and video presentations, as appropriate. If clinically indicated, the healthcare professional or nurse will make a referral to a lactation consultant or specialist for additional education or assistance. (II-1, II-2, III)

10. Breastfeeding mothers will be instructed about:

a. Proper positioning and latch on
b. Nutritive suckling and swallowing
c. Milk production and release
d. Frequency of feeding/feeding cues
e. Hand expression of breast milk and use of a pump if indicated
f. How to assess if infant is adequately nourished
g. Reasons for contacting the healthcare professional

These skills will be taught to primiparous and multiparous women, provided in written form (Eidelman, Hoffmann, & Kaitz, 1993), and reviewed before the mother goes home. (II-1, II-2, III)

11. Parents will be taught that breastfeeding infants, including cesarean-birth babies, should be put to breast at least 8 to 12 times each 24 hours, with some infants needing to be fed more frequently. Infant feeding cues (e.g., increased alertness or activity, mouthing, or rooting) will be used as indicators of the baby´s readiness for feeding. Breastfeeding babies will be breastfed at night. (II-1, II-2, III)

12. Time limits for breastfeeding on each side will be avoided. Infants can be offered both breasts at each feeding but may be interested in feeding only on one side at a feeding during the early days. (II-1, II-2, III)

13. No supplemental water, glucose water, or formula will be given unless specifically ordered by a healthcare professional (e.g., physician, certified nurse midwife, or nurse practitioner) or by the mother´s documented and informed request. Prior to non-medically indicated supplementation, mothers will be informed of the risks of supplementing. The supplement should be fed to the baby by cup if possible and will be no more than 10 to 15 mL (per feeding) in a term baby (during the first 1 to 2 days of life). Alternative feeding methods such as syringe or spoon feeding may also be used; however, these methods have not been shown to be effective in preserving breastfeeding. Bottles will not be placed in a breastfeeding infant´s bassinet (Howard et al., 2003; Howard et al., 1999; Marinelli, Burke, & Dodd, 2001). (II-1, II-2)

14. This institution does not give group instruction in the use of formula. Those parents who, after appropriate counseling, choose to formula feed their infants will be provided individual instruction.

15. Pacifiers will not be given to normal full-term breastfeeding infants. The pacifier guidelines at "name of institution" state that

preterm infants in the Neonatal Intensive Care or Special Care Unit or infants with specific medical conditions (e.g., neonatal abstinence syndrome) may be given pacifiers for non-nutritive sucking. Newborns undergoing painful procedures (e.g., circumcision) may be given a pacifier as a method of pain management during the procedure. The infant will not return to the mother with the pacifier. "Name of institution" encourages "pain-free newborn care," which may include breastfeeding during the heel stick procedure for the newborn metabolic screening tests (Gray et al., 2002). (I)

16. Routine blood glucose monitoring of full-term healthy appropriate-for-gestational age infants is not indicated. Assessment for clinical signs of hypoglycemia and dehydration will be ongoing (Wight, Marinelli, & Academy of Breastfeeding Medicine Clinical Protocol Committee, 2006). (I)

17. Antilactation drugs will not be given to any postpartum mother. (I)

18. Routine use of nipple creams, ointments, or other topical preparations will be avoided unless such therapy has been indicated for a dermatologic problem. Mothers with sore nipples will be observed for latch-on techniques and will be instructed to apply expressed colostrum or breast milk to the areola/nipple after each feeding. (III)

19. Nipple shields or bottle nipples will not be routinely used to cover a mother´s nipples, to treat latch-on problems, or to prevent or manage sore or cracked nipples or used when a mother has flat or inverted nipples. Nipple shields will be used only in conjunction with a lactation consultation and after other attempts to correct the difficulty have failed. (III)

20. After 24 hours of life, if the infant has not latched on or fed effectively, the mother will be instructed to begin to massage her breasts and hand express colostrum into the baby´s mouth during feeding attempts. Skin-to-skin contact will be encouraged. Parents will be instructed to watch closely for feeding cues and whenever these are observed to awaken and feed the infant. If the baby continues to feed poorly, hand expression by the mother or a double set-up electric breast pump will be initiated and maintained approximately every 3 hours or a minimum of eight times per day. Any expressed colostrum or mother´s milk will be fed to the baby by an alternative method. The mother will be reminded that she may not obtain much milk or even any milk the first few times she expresses her breasts. Until the mother´s milk is available, a collaborative decision should be made among the mother, nurse, and healthcare professional (e.g., physician/nurse practitioner/certified nurse midwife) regarding the need to supplement the baby. Each day the responsible healthcare professional will be consulted regarding the volume and type of the supplement. Pacifiers will be avoided. In cases of problem feeding, the lactation consultant or specialist will be consulted (Academy of Breastfeeding Medicine Protocol Committee, "ABM clinical protocol #3," 2009). (I, III)

21. If the baby is still not latching on well or feeding well when discharged to home, the feeding/expression/supplementing plan will be reviewed in addition to routine breastfeeding instructions. A follow-up visit or contact will be scheduled within 24 hours. Depending on the clinical situation it may be appropriate to delay discharge of the couplet to provide further breastfeeding intervention, support, and education. (III)

22. All babies should be seen for follow-up within the first few days postpartum. This visit should be with a physician (pediatrician or family physician) or other qualified health care practitioner for a formal evaluation of breastfeeding performance, a weight check, assessment of jaundice and age appropriate elimination: (a) for infants discharged at less than 2 days of age (<48 hours), follow-up at 2 to 4 days of age; (b) for infants discharged between 48 and 72 hours, follow-up at 4 to 5 days of age. Infants discharged after 5 to 6 days may be seen 1 week later.

23. Mothers who are separated from their sick or premature infants will be

a. Instructed on how to use skilled hand expression or the double set up electric breast pump. Instructions will include expression at least eight times per day or approximately every 3 hours for 15 minutes (or until milk flow stops, whichever is greater) around the clock and the importance of not missing an expression session during the night (III)

b. Encouraged to breastfeed on demand as soon as the infant´s condition permits (III)

c. Taught proper storage and labeling of human milk (III)

d. Assisted in learning skilled hand expression or obtaining a double set-up electric breast pump prior to going home (III)

24. Before leaving the hospital (Academy of Breastfeeding Medicine Clinical Protocol Committee, 2007), breastfeeding mothers should be able to:

a. Position the baby correctly at the breast with no pain during the feeding

b. Latch the baby to breast properly

c. State when the baby is swallowing milk

d. State that the baby should be nursed a minimum of eight to 12 times a day until satiety, with some infants needing to be fed more frequently

e. State age-appropriate elimination patterns (at least six urinations per day and three to four stools per day by the fourth day of life)

f. List indications for calling a healthcare professional

g. Manually express milk from their breasts (III)

25. Prior to going home, mothers will be given the names and telephone numbers of community resources to contact for help with breastfeeding, including (the support group or resource recommended by "name of institution").

26. "Name of institution" does not accept free formula or free breast milk substitutes. Nursery or Neonatal Intensive Care Unit discharge bags offered to all mothers will not contain infant formula, coupons for formula, logos of formula companies, or literature with formula company logos.
27. "Name of institution" health professionals will attend educational sessions on lactation management and breastfeeding promotion to ensure that correct, current, and consistent information is provided to all mothers wishing to breastfeed (American Academy of Pediatrics, American Academy of Obstetricians and Gynecologists, 2006).
Contraindications:
Breastfeeding is contraindicated in the following situations:
- Mothers who are human immunodeficiency virus (HIV)-positive in locations where artificial feeding is acceptable, feasible, affordable, sustainable, and safe (I)
- Mothers currently using illicit drugs (e.g., cocaine, heroin) unless specifically approved by the infant´s healthcare provider on a case-by-case basis (I)
- Mothers taking certain medications. Most prescribed and over-the-counter drugs are safe for the breastfeeding infant. Some medications may make it necessary to interrupt breastfeeding, such as radioactive isotopes, antimetabolites, cancer chemotherapy, some psychotropic medications and a small number of other medications. (III)
- Mothers with active, untreated tuberculosis. A mother can express her milk until she is no longer contagious. (I)
- Infants with galactosemia (I)
- Mothers with active herpetic lesions on her breast(s). Breastfeeding can be recommended on the unaffected breast. (The Infectious Disease Service will be consulted for problematic infectious disease issues.) (I)
- Mothers with onset of varicella within 5 days before or up to 48 hours after delivery, until they are no longer infectious (I)
- Mothers with human T-cell lymphotropic virus type I or type II (I)

1c.17 Clinical Practice Guideline Citation: • Philipp BL, Academy of Breastfeeding Medicine Protocol Committee. ABM clinical protocol #7: model breastfeeding policy (revision 2010). Breastfeed Med 2010 Aug;5(4):173-7.

1c.18 National Guideline Clearinghouse or other URL:
http://www.guideline.gov/content.aspx?id=24013&search=breastfeeding+policy

1c.19 Grading of Strength of Guideline Recommendation. Has the recommendation been graded? Yes

1c.20 If guideline recommendation graded, identify the entity that graded the evidence including balance of representation and any disclosures regarding bias: Academy of Breastfeeding Medicine Protocol Committee

1c.21 System Used for Grading the Strength of Guideline Recommendation: Other

1c.22 If other, identify and describe the grading scale with definitions: The system for categorizing recommendations in this guideline is as follows:

Levels of Evidence
I Evidence obtained from at least one properly randomized controlled trial
II-1 Evidence obtained from well-designed controlled trials without randomization
II-2 Evidence obtained from well-designed cohort or case-control analytic studies, preferably from more than one center or research group
II-3 Evidence obtained from multiple time series with or without the intervention. Dramatic results in uncontrolled experiments (such as the results of the introduction of penicillin treatment in the 1940s) could also be regarded as this type of evidence.
III Opinions of respected authorities, based on clinical experience, descriptive studies and case reports; or reports of expert committees

1c.23 Grade Assigned to the Recommendation: Grading varies from I to III

1c.24 Rationale for Using this Guideline Over Others: This policy is based on recommendations from the most recent breastfeeding policy statements published by the Office on Women´s Health of the U.S. Department of Health and Human Services, the American Academy of Pediatrics, the American College of Obstetricians and Gynecologists, the American Academy of Family Physicians, the World Health Organization (WHO), the Academy of Breastfeeding Medicine, and the UNICEF/WHO evidence-based "Ten Steps to Successful Breastfeeding."
The recommendations were based primarily on a comprehensive review of the existing literature. In cases where the literature does not

appear conclusive, recommendations were based on the consensus opinion of the group of experts.
Based on the NQF descriptions for rating the evidence, what was the <u>developer's assessment</u> of the quantity, quality, and consistency of the body of evidence? **1c.25** Quantity: High **1c.26** Quality: High**1c.27** Consistency**:** High

1. Evidence, Performance Gap, Priority – Importance to Measure and Report

Extent to which the specific measure focus is evidence-based, important to making significant gains in healthcare quality, and improving health outcomes for a specific high-priority (high-impact) aspect of healthcare where there is variation in or overall less-than-optimal performance. ***Measures must be judged to meet all subcriteria to pass this criterion and be evaluated against the remaining criteria.***

1a. Evidence to Support the Measure Focus – See attached Evidence Submission Form
0480_Evidence_MSF5.0_Data-635787043120686011.doc

1b. Performance Gap

Demonstration of quality problems and opportunity for improvement, i.e., data demonstrating:

- considerable variation, or overall less-than-optimal performance, in the quality of care across providers; and/or
- disparities in care across population groups.

1b.1. **Briefly explain the rationale for this measure** *(e.g., the benefits or improvements in quality envisioned by use of this measure)*
Increasing the number of newborns who are exclusively fed breast milk for the first six months of life remains a major goal of the WHO, DHHS, AAP and ACOG. Guidelines for the promotion of breast milk feeding are available from the CDC to assist hospitals in establishing successful interventions to improve exclusive breast milk feeding rates in newborns. Breast milk feeding results in numerous health benefits for both mother and newborn. Breastfeeding is associated with decreased risk for many early-life diseases and conditions, including otitis media, respiratory tract infections, atopic dermatitis, gastroenteritis, type 2 diabetes, sudden infant death syndrome, and obesity. Breastfeeding also is associated with health benefits to women, including decreased risk for type 2 diabetes, ovarian cancer, and breast cancer
The measure will assist health care organizations (HCOs) to track evidence of an increase in the number of newborns who were exclusively fed breast milk during the birth hospitalization.

1b.2. **Provide performance scores on the measure as specified (current and over time) at the specified level of analysis.** *(This is required for endorsement maintenance. Include mean, std dev, min, max, interquartile range, scores by decile. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included). This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*
Rates for exclusive breast milk feeding remain below 50% for over half of the hospitals reporting data. A goal of 70% should be achievable based on the Joint Commission's analysis of the data. The Perinatal Care (PC) core measures were added as a new core measure set in 2010 for hospitals to select in order to meet their ORYX performance measurement requirement for Joint Commission accreditation purposes at that time, approximately 165 hospitals reported the data with an average measure rate of 40.9% (n=54,630 patients). In January 2014, The Joint Commission required mandatory reporting of the PC measure set for all accredited hospitals with 1100 births or more annually. 1386 hospitals reported the data with an average rate of 49.4% (n=728,157 patients). It is important to note that a performance gap of 48% exists for the 10th percentile of hospitals performing at 22.0% (if 70% is considered goal performance) and that the aggregate rate dropped by 4.2% from the 2013 rate of 53.6%. Only the 90th percentile hospitals (74.3%) are performing above the goal of 70%. The threshold for mandatory reporting was recently lowered to 300 births annually effective January 2016. The new reporting requirement will now capture approximately 80% of all accredited birthing hospitals. As a result, the rates may decrease with the addition of approximately 821 more hospitals reporting data. Below is the specified level of analysis for PC-01 beginning with discharges April 1, 2010 through December 31, 2014.
2Q 2010: 54,630 denominator cases; 22,346 numerator cases; 165 hospitals; 40.9% national aggregate rate; 0.46132 mean of hospital rates; 0.22274 standard deviation; 75.3% 90th percentile rate; 62.2% 75th percentile rate/upper quartile; 45.9% 50th percentile rate/median rate; 29.4% 25th percentile rate/lower quartile; and 15.5% 10th percentile rate.
CY 2011: 69,613 denominator cases; 31,999 numerator cases; 166 hospitals; 45.9% national aggregate rate; 0.49335 mean of hospital rates; 0.21348 standard deviation; 76.2% 90th percentile rate; 62.8% 75th percentile rate/upper quartile; 50.6% 50th percentile rate/median rate; 34.8% 25th percentile rate/lower quartile; and 19.9% 10th percentile rate.
CY 2012: 76,952 denominator cases; 39,337 numerator cases; 170 hospitals; 51.1% national aggregate rate; 0.55872 mean of hospital rates; 0.20359 standard deviation; 80.7% 90th percentile rate; 72.2% 75th percentile rate/upper quartile; 56.7% 50th percentile rate/median rate; 41.7% 25th percentile rate/lower quartile; and 27.1% 10th percentile rate.
CY 2013: 91,011 denominator cases; 48,758 numerator cases; 197 hospitals; 53.6% national aggregate rate; 0.5632 mean of hospital rates; 0.19707 standard deviation; 79.3% 90th percentile rate; 72.1%% 75th percentile rate/upper quartile; 57.3% 50th percentile rate/median rate; 42.3% 25th percentile rate/lower quartile; and 28.4% 10th percentile rate.
CY 2014: 728,157 denominator cases; 359,633 numerator cases; 1386 hospitals; 49.4% national aggregate rate; 0.48724 mean of

hospital rates; 0.19475 standard deviation; 74.3% 90th percentile rate; 62.4% 75th percentile rate/upper quartile; 49.1% 50th percentile rate/median rate; 35.1% 25th percentile rate/lower quartile; and 22.0% 10th percentile rate.

1b.3. **If no or limited performance data on the measure as specified is reported in 1b2, then provide a summary of data from the literature that indicates opportunity for improvement or overall less than optimal performance on the specific focus of measurement.**
Not Applicable

1b.4. **Provide disparities data from the measure as specified (current and over time) by population group, e.g., by race/ethnicity, gender, age, insurance status, socioeconomic status, and/or disability.** *(This is required for endorsement maintenance. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities include.) This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*
A study was conducted by Petrova et al. (2007) to identify the association between the in-hospital feeding pattern and the infant´s post discharge feeding modality during the first month of life in a culturally diverse population of women. Demographic, clinical, and feeding practice data were collected from the medical charts and interviews of mothers conducted in the first month after singleton delivery of healthy term newborns. Among the 307 mothers who completed the study, exclusive in-hospital breast milk feeding was reported by 54.2% of white, 38.7% of black, 54.0% of asian, and 44.7% of hispanic ($p = 0.063$), and among these, only 55.6%, 50.0%, 58.9%, and 19.1%, respectively, maintained exclusive breast milk feeding during the first postpartum month ($p < 0.02$). The rate of exclusive breast milk feeding at the end of the first month was 10.5%, 15.8%, 20.7%, and 3.9%, respectively, for the white, black, asian, and hispanic mothers whose infants received partial or no breastfeeding in-hospital.

Overall, the logistic regression analysis showed significant association between initiation of exclusive breast milk feeding in-hospital and exclusive breast milk feeding at the end of the first month (odds ratio 7.2 and 95% confidence interval 4.0, 12.6). In conclusion, it showed a larger decline in the continuation of exclusive breast milk feeding and the lowest rate of exclusive breast milk feeding at 1 month in the hispanic mothers. Irrespective of race/ethnicity, mothers who practice exclusive breast milk feeding in-hospital are more likely to exclusively fed breast milk throughout the neonatal period.

According to the CDC, from 2000-2004 the rates of exclusive breastfeeding were significantly lower among black infants (compared with white infants) and infants born to unmarried mothers (compared with married mothers). Additionally, older age, urban residence, higher education, and higher income of mothers all were positively associated with exclusive breast milk feeding (CDC, 2007). Hawkins et al. (2015) noted continued disparities among mothers with lower education based on Pregnancy Risk Assessment Monitoring System (PRAMS) data collected from 1999 to 2009.

1b.5. **If no or limited data on disparities from the measure as specified is reported in 1b4, then provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement. Include citations.**

- Centers for Disease Control and Prevention (CDC). (2007). Breastfeeding trends and updated national health objectives for exclusive breastfeeding-United States, birth years 2000-2004. [Journal Article] MMWR - Morbidity & Mortality Weekly Report. 56(30):760-3.
- Hawkins, S., Stern, A., Baum, C. & Gillman, M. (2015). Evaluating the impact of the baby-friendly hospital initiative on breast-feeding rates: a multi-state analysis. Public Health Nutr.18(2):189-97.
- Petrova, A., Hegyi, T., Mehta, R. (2007). Maternal race/ethnicity and one-month exclusive breastfeeding in association with the in-hospital feeding modality. Breastfeeding Medicine: The Official Journal of the Academy of Breastfeeding Medicine. 2(2):92-8.

1c. High Priority (previously referred to as High Impact)
The measure addresses:
- a specific national health goal/priority identified by DHHS or the National Priorities Partnership convened by NQF; OR
- a demonstrated high-priority (high-impact) aspect of healthcare (e.g., affects large numbers of patients and/or has a substantial impact for a smaller population; leading cause of morbidity/mortality; high resource use (current and/or future); severity of illness; and severity of patient/societal consequences of poor quality).

1c.1. **Demonstrated high priority aspect of healthcare**
Affects large numbers, Patient/societal consequences of poor quality
1c.2. **If Other:**

1c.3. **Provide epidemiologic or resource use data that demonstrates the measure addresses a high priority aspect of healthcare. List citations in 1c.4.**

Exclusive breast milk feeding for the first 6 months of neonatal life has long been the expressed goal of World Health Organization (WHO), US Department of Health and Human Services (DHHS), American Academy of Pediatrics (AAP) and American College of Obstetricians and Gynecologists (ACOG). ACOG reiterated its position in a Committee Opinion on breastfeeding (ACOG, 2007). Additionally, a Cochrane review of two randomized control trials and 18 other studies substantiates the benefits of exclusive breast milk feeding for the first six months of life (Kramer et al., 2002).

Breastfeeding is associated with decreased risk for many early-life diseases and conditions, including otitis media, respiratory tract infections, atopic dermatitis, gastroenteritis, type 2 diabetes, sudden infant death syndrome, and obesity. Breastfeeding also is associated with health benefits to women, including decreased risk for type 2 diabetes, ovarian cancer, and breast cancer (Ip et al., 2007). Exclusive breastfeeding is defined as a newborn receiving only breast milk and no other liquids or solids except for drops or syrups consisting of vitamins, minerals, or medicines (WHO, 1991).

In 2007, Healthy People 2010 objectives for breastfeeding initiation and duration included two new objectives on exclusive breastfeeding to increase the proportion of mothers who exclusively breastfeed their infants through age 3 months to 60% and through age 6 months to 25% [objectives 16-19d and 16-19e] (DHHS, 2000). The Healthy People 2020 objectives for exclusive breastfeeding were continued through age 3 months with a goal of 46.2% and age 6 months with a goal of 25.5% [objectives MICH-21.4 and MICH-21.5]. Also included is the related objective MICH-24: increase the proportion of live births that occur in facilities that provide recommended care for lactating mothers and their babies (DHHS, 2010).

The Centers for Disease Control and Prevention (CDC) developed a Guide to Breastfeeding Interventions in 2005 for the promotion and support of breastfeeding based on detailed input from the spectrum of breastfeeding experts which can be used to help hospitals achieve the Healthy People 2020 objective MICH-24. Institutional changes i.e., attaining Baby Friendly Hospital Initiative status, individual interventions including increased rooming-in of mothers and newborns, early skin to skin contact and discontinuing policies that are not evidence based have been shown to increase breastfeeding initiation and duration rates as well (Shealy et al., 2007). According to the CDC (2011), mothers who want to breastfeed who do not receive hospital support will stop early. The CDC encourages hospitals to partner with Baby-Friendly hospitals to learn how to improve maternity care, use the CDC's Maternity Practice in Infant Nutrition and Care (mPINC) survey data to prioritize changes to improve maternity care practices and stop distributing formula samples and give-aways to breastfeeding mothers.

1c.4. Citations for data demonstrating high priority provided in 1a.3

• American College of Obstetricians and Gynecologists (ACOG). (Feb. 2007). Committee on Obstetric Practice and Committee on Health Care for Underserved Women. Breastfeeding: Maternal and Infant Aspects. ACOG Committee Opinion 361.
• Brown, P., Kaiser, K. & Nailon,R. (2014). Integrating quality improvement and translational research models to increase exclusive breastfeeding. Journal of Obstetric, Gynecologic, and Neonatal Nursing : JOGNN / NAACOG. 43:5,545-553.
• Centers for Disease Control and Prevention (CDC). (2011). Hospital support for breastfeeding: Preventing obesity begins in hospitals. CDC Vital Signs, Retrieved September 26, 2011 at: http://www.cdc.gov/VitalSigns/pdf/2011-08-vitalsigns.pdf
• Chantry, C., Eglash, A. & Labbok, M.(2015). ABM position on breastfeeding. Breastfeeding Medicine. 10(9): 407-411.
• Chantry, C., Dewey,K., Peerson, J., Wagner, E. & Nommsen-Rivers,L. (2014). In-hospital formula use increases early breastfeeding cessation among first-time mothers intending to exclusively breastfeed.The Journal of Pediatrics. 164:6, 1339-45.e5.
• Dias de Oliveira, L., Justo Giugliani,E., Cordova do Espirito Santo,L. & Meirelles Nunes,L. (2014). Counselling sessions increased duration of exclusive breastfeeding: a randomized clinical trial with adolescent mothers and grandmothers. Nutrition Journal. 13, 73-2891-13-73.
• Ip S, Chung M, Raman G, et al. Breastfeeding and maternal and infant health outcomes in developed countries. Rockville, MD: US Department of Health and Human Services; 2007. Retrieved on February 29, 2016 at: http://archive.ahrq.gov/downloads/pub/evidence/pdf/brfout/brfout.pdf
• Kramer, M.S. & Kakuma, R. (2002).Optimal duration of exclusive breastfeeding. [107 refs] Cochrane Database of Systematic Reviews. (1):CD003517.
• Linares, A., Rayens, M., Gomez, M, Gokun,Y. & Dignan,M. (2014). Intention to breastfeed as a predictor of initiation of exclusive breastfeeding in hispanic women. Journal of Immigrant and Minority Health / Center for Minority Public Health. DOI: 10.1007/s10903-014-0049-0.
• Mueffelmann, R., Racine, E., Warren-Findlow,J. & Coffman,M. (2014). Perceived infant feeding preferences of significant family members and mother´s intention to exclusively breastfeed. Journal of Human Lactation : Official Journal of International Lactation Consultant Association, doi: 10.1177/0890334414553941.
• Perrine C.,et al., (2015). How do hospitals measure up in supporting breast-feeding?. MMWR Morb Mortal Wkly Rep. 64:1112
• Ramakrishnan, R.,Oberg, C. & Kirby,R. (2014).The association between maternal perception of obstetric and pediatric care providers´ attitudes and exclusive breastfeeding outcomes. Journal of Human Lactation : Official Journal of International Lactation Consultant Association.3:1,80-87.

• Shealy, K.R., Li, R., Benton-Davis, S., & Grummer-Strawn, L.M. (2005).The CDC guide to breastfeeding interventions. Atlanta, GA: US Department of Health and Human Services, CDC. Available at: http://www.cdc.gov/breastfeeding/pdf/breastfeeding_interventions.pdf
• US Department of Health and Human Services (DHHS). (2000). Healthy People 2010. Washington, DC. Retrieved on September 26, 2011 at: http://www.healthypeople.gov/2010
• US Department of Health and Human Services (DHHS). (2010). Healthy People 2020. Washington, DC. Retrieved on September 26, 2011 at: http://www.healthypeople.gov/2020
• World Health Organization (WHO). Indicators for assessing breastfeeding practices. Geneva, Switzerland: World Health Organization; 1991. Retrieved on September 27, 2011 at: http://www.who.int/child-adolescent-health/new_publications/nutrition/who_cdd_ser_91.14.pdf

1c.5. **If a PRO-PM (e.g. HRQoL/functional status, symptom/burden, experience with care, health-related behaviors), provide evidence that the target population values the measured PRO and finds it meaningful.** *(Describe how and from whom their input was obtained.)*
Not Applicable

2. Reliability and Validity—Scientific Acceptability of Measure Properties

Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. ***Measures must be judged to meet the subcriteria for both reliability and validity to pass this criterion and be evaluated against the remaining criteria.***

2a.1. Specifications The measure is well defined and precisely specified so it can be implemented consistently within and across organizations and allows for comparability. eMeasures should be specified in the Health Quality Measures Format (HQMF) and the Quality Data Model (QDM).

De.5. **Subject/Topic Area** *(check all the areas that apply):*
Perinatal and Reproductive Health, Perinatal and Reproductive Health : Newborn

De.6. **Cross Cutting Areas** *(check all the areas that apply):*
Patient and Family Engagement

S.1. **Measure-specific Web Page** *(Provide a URL link to a web page specific for this measure that contains current detailed specifications including code lists, risk model details, and supplemental materials. Do not enter a URL linking to a home page or to general information.)*
http://manual.jointcommission.org/releases/TJC2015B2/

S.2a. **If this is an eMeasure**, HQMF specifications must be attached. Attach the zipped output from the eMeasure authoring tool (MAT) - if the MAT was not used, contact staff. (Use the specification fields in this online form for the plain-language description of the specifications)
This is not an eMeasure **Attachment:**

S.2b. **Data Dictionary, Code Table, or Value Sets** *(and risk model codes and coefficients when applicable) must be attached. (Excel or csv file in the suggested format preferred - if not, contact staff)*
Attachment **Attachment:** PC05_ICD_Code_Tables.xlsx

S.3. **For endorsement maintenance**, please briefly describe any changes to the measure specifications since last endorsement date and explain the reasons.

- All ICD-9-CM diagnosis codes and ICD-9-CM procedure codes were converted to ICD-10-CM diagnosis and ICD-10-PCS procedure codes throughout the measure specifications. The Department of Health and Human Services (HHS) mandated that all entities covered by the Health Insurance Portability and Accountability Act (HIPAA) must transition to a new set of codes for electronic health care transactions on October 1, 2015.
- The sub-measure PC-05a: Exclusive Breast Milk Feeding Considering Mother's Initial Feeding Plan was retired, based on feedback from key stakeholders – including health care organizations, The Centers for Disease Control and Prevention (CDC), the American College of Obstetricians and Gynecologists (ACOG), the American Academy of Pediatrics (AAP) and the Association of Women's Health, Obstetric & Neonatal Nurses (AWHONN). The sub-measure greatly increased the burden of data abstraction, did not follow normal workflow patterns and impeded improved performance for PC-05.
- The denominator data element Reason for Not Exclusively Feeding Breast Milk was removed as a denominator exclusion, since this concept cannot be modeled in the eCQM version of the measure and removal greatly reduces the burden of data abstraction.
- The initial patient population now identifies eligible patients with ICD-10-CM-Principal Diagnosis Code for newborn born at hospital and data elements Point of Origin for Admission or Visit and Admission Type were removed, since updates to the latter data elements are not available in the public domain.
- Late preterm newborns are now excluded from the denominator population with the new data element Term Newborn instead of using diagnosis codes for premature newborns found on Table 11.23 in Appendix A. This change was made because prematurity codes were not being routinely assigned unless birth weight codes are also assigned.

S.4. **Numerator Statement** *(Brief, narrative description of the measure focus or what is being measured about the target population, i.e., cases from the target population with the target process, condition, event, or outcome)*
IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
Newborns that were fed breast milk only since birth

S.5. **Time Period for Data** *(What is the time period in which data will be aggregated for the measure, e.g., 12 mo, 3 years, look back*

to August for flu vaccination? Note if there are different time periods for the numerator and denominator.)
Episode of care

S.6. **Numerator Details** *(All information required to identify and calculate the cases from the target population with the target process, condition, event, or outcome such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
IF an OUTCOME MEASURE, describe how the observed outcome is identified/counted. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
One data element is used to calculate the numerator:

1. Exclusive Breast Milk Feeding - Documentation that the newborn was exclusively fed breast milk during the entire hospitalization. Allowable Values: Yes or No/UTD. Cases are eligible for the numerator when allowable value = yes. Updates available at: http://manual.jointcommission.org/releases/TJC2015B2/

S.7. **Denominator Statement** *(Brief, narrative description of the target population being measured)*
Single term liveborn newborns discharged alive from the hospital with ICD-10-CM Principal Diagnosis Code for single liveborn newborn as defined in Appendix A, Table 11.20.1 available at: http://manual.jointcommission.org/releases/TJC2015B2/

S.8. **Target Population Category** *(Check all the populations for which the measure is specified and tested if any):*
Maternal Health

S.9. **Denominator Details** *(All information required to identify and calculate the target population/denominator such as definitions, specific data collection items/responses , code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
Eleven data elements are used to calculate the denominator:

1. Admission Date – The month, day and year of admission to acute inpatient care.
2. Admission to NICU - Documentation that the newborn was admitted to the Neonatal Intensive Care Unit (NICU) at this hospital any time during the hospitalization. Allowable values: Yes or No/UTD
3. Birthdate - The month, day and year the patient was born.
4. Clinical Trial - Documentation that during this hospital stay the patient was enrolled in a clinical trial in which patients who are newborns were being studied. Allowable values: Yes or No/UTD
5. Discharge Date – The month day and year the patient was discharged from acute care, left against medical advice or expired during the stay.
6. Discharge Disposition - The place or setting to which the patient was discharged.
7. ICD-10-CM Other Diagnosis Codes - The International Classification of Diseases, Tenth Revision, Clinical Modification codes associated with the secondary diagnoses for this hospitalization.
8. ICD-10-PCS Other Procedure Codes - The International Classification of Diseases, Tenth Revision, Procedure Coding System code that identifies significant procedures performed other than the principal procedure during this hospitalization.
9. ICD-10-CM Principal Diagnosis Code - The International Classification of Diseases, Tenth Revision, Clinical Modification code associated with the diagnosis established after study to be chiefly responsible for occasioning the admission of the patient for this hospitalization.
10. ICD-10-CM Principal Procedure Code - The International Classification of Diseases, Tenth Revision, Procedure Coding System code that identifies the principal procedure performed during this hospitalization. The principal procedure is the procedure performed for definitive treatment rather than diagnostic or exploratory purposes, or which is necessary to take care of a complication.
11. Term Newborn - Documentation that the newborn was at term or >= 37 completed weeks of gestation at the time of birth. Allowable values: Yes or No/UTD
Updates available at: http://manual.jointcommission.org/releases/TJC2015B2/

S.10. **Denominator Exclusions** *(Brief narrative description of exclusions from the target population)*
- Admitted to the Neonatal Intensive Care Unit (NICU) at this hospital during the hospitalization
- ICD-10-CM Other Diagnosis Codes for galactosemia as defined in Appendix A, Table 11.21
- ICD-10-PCS Principal Procedure Code or ICD-10-PCS Other Procedure Codes for parenteral infusion as defined in Appendix A, Table 11.22
- Experienced death

• Length of Stay >120 days
• Enrolled in clinical trials
• Patients transferred to another hospital
• Patients who are not term or with < 37 weeks gestation completed

S.11. **Denominator Exclusion Details** *(All information required to identify and calculate exclusions from the denominator such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
• The data element Admission to NICU is used to determine if the patient was admitted to the NICU.
• Patients with ICD-10-CM Other Diagnosis Codes for galactosemia are excluded.
• Patients with ICD-10-PCS Principal Procedure Code or ICD-10-PMS Other Procedure Codes for parenteral infusion are excluded.
• Length of stay (LOS) in days is equal to the Discharge Date minus the Admission Date. If the LOS is greater than 120 days the patient is excluded.
• Patients are excluded if "Yes" is selected for Clinical Trial.
• The data element Discharge Disposition is used to determine if the patient was transferred to another hospital or expired.
• The data element Term Newborn is used to determine if the patient was not term or < 37 completed weeks of gestation.

S.12. **Stratification Details/Variables** *(All information required to stratify the measure results including the stratification variables, definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format with at S.2b)*
Not Applicable

S.13. **Risk Adjustment Type** (Select type. Provide specifications for risk stratification in S.12 and for statistical model in S.14-15)
No risk adjustment or risk stratification
If other:

S.14. **Identify the statistical risk model method and variables** *(Name the statistical method - e.g., logistic regression and list all the risk factor variables. Note - risk model development and testing should be addressed with measure testing under Scientific Acceptability)*
Not Applicable

S.15. **Detailed risk model specifications** *(must be in attached data dictionary/code list Excel or csv file. Also indicate if available at measure-specific URL identified in S.1.)*
Note: Risk model details (including coefficients, equations, codes with descriptors, definitions), should be provided on a separate worksheet in the suggested format in the Excel or csv file with data dictionary/code lists at S.2b.

S.15a. **Detailed risk model specifications** (*if not provided in excel or csv file at S.2b*)
Not Applicable

S.16. **Type of score:**
Rate/proportion
If other:

S.17. **Interpretation of Score** *(Classifies interpretation of score according to whether better quality is associated with a higher score, a lower score, a score falling within a defined interval, or a passing score)*
Better quality = Higher score

S.18. **Calculation Algorithm/Measure Logic** *(Describe the calculation of the measure score as an ordered sequence of steps including identifying the target population; exclusions; cases meeting the target process, condition, event, or outcome; aggregating data; risk adjustment; etc.)*
1. Start processing. Run cases that are included in the PC-Newborn Initial Patient Newborns with Breast Feeding and pass the edits defined in the Transmission Data Processing Flow: Clinical through this measure.

2. Check Discharge Disposition
a. If Discharge Status equals 4, 6, the case will proceed to a Measure Category Assignment of B and will not be in the measure population. Stop processing.
b. If Discharge Status equals 1, 2, 3, 5, 7, 8, continue processing and proceed to Clinical Trial.

3. Check Clinical Trial
a. If Clinical Trial is missing, the case will proceed to a Measure Category Assignment of X and will be rejected. Stop processing.
b. If Clinical Trial equals Yes, the case will proceed to a Measure Category Assignment of B and will not be in the measure population. Stop processing.
c. If Clinical Trial equals No, continue processing and proceed to Term Newborn.

4. Check Term Newborn
a. If Term Newborn is missing, the case will proceed to a Measure Category Assignment of X and will be rejected. Stop processing.
b. If Term Newborn equals Yes, the case will proceed to a Measure Category Assignment of B and will not be in the measure population. Stop processing.
c. If Term Newborn equals No, continue processing and proceed to Admission to NICU.

5. Check Admission to NICU
a. If Admission to NICU is missing, the case will proceed to a Measure Category Assignment of X and will be rejected. Stop processing.
b. If Admission to NICU equals Yes, the case will proceed to a Measure Category Assignment of B and will not be in the measure population. Stop processing.
c. If Admission to NICU equals No, continue processing and proceed to Exclusive Breast Milk Feeding.

6. Check Exclusive Breast Milk Feeding
a. If Exclusive Breast Milk Feeding is missing, the case will proceed to a Measure Category Assignment of X and will be rejected. Stop processing.
b. If Exclusive Breast Milk Feeding equals Yes, the case will proceed to a Measure Category Assignment of E and will be in the Numerator Population. Stop processing.
c. If Exclusive Breast Milk Feeding equals No, the case will proceed to a Measure Category Assignment of D and will be in the Measure Population. Stop processing.

S.19. **Calculation Algorithm/Measure Logic Diagram URL or Attachment** *(You also may provide a diagram of the Calculation Algorithm/Measure Logic described above at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
Available at measure-specific web page URL identified in S.1

S.20. **Sampling** *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*
IF a PRO-PM, identify whether (and how) proxy responses are allowed.
The initial patient population includes patients with age at admission (Admission Date – Birthdate) = 2 days, Length of Stay (Discharge Date - Admission Date) = 120 days, an ICD-10-CM Principal or Other Diagnosis Code as defined in Appendix A, Table 11.20.1, NO ICD-10-CM Principal or Other Diagnosis Code as defined in Appendix A, Table 11.21 and NO ICD-10-PCS-Principal or Other Procedure Code as defined in Appendix A, Table 11.22 and NO ICD-10-CM Principal or Other Diagnosis Code as defined in Appendix A are included in this subpopulation and are eligible to be sampled. The sample is taken randomly as follows for a monthly sample:
- Average monthly Initial Patient Population >= 181 results in a minimum random sample size of 37.
- Average monthly Initial Patient Population 46 – 180 results in a minimum random sample size of 20% of the population size.
- Average monthly Initial Patient Population 9 – 45 results in a minimum random sample size of 9.
- Average monthly Initial Patient Population <9 results in no sampling; 100% Initial Patient Population required

S.21. **Survey/Patient-reported data** *(If measure is based on a survey, provide instructions for conducting the survey and guidance on minimum response rate.)*
IF a PRO-PM, specify calculation of response rates to be reported with performance measure results.
Not Applicable

S.22. **Missing data** (specify how missing data are handled, e.g., imputation, delete case.)
Required for Composites and PRO-PMs.
Any file with missing data will result in a measure category assignment of X and rejection of the file from the warehouse, unless the data are not used to process the measure. If the data are used to process the measure and have been reported by the abstractor as No/UTD, the case will result in a measure category assignment of E (i.e., failed measure, not rejected.

S.23. **Data Source** *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*
If other, please describe in S.24.
Electronic Clinical Data, Paper Medical Records

S.24. **Data Source or Collection Instrument** *(Identify the specific data source/data collection instrument e.g. name of database, clinical registry, collection instrument, etc.)*
IF a PRO-PM, identify the specific PROM(s); and standard methods, modes, and languages of administration.
Each data element in the data dictionary includes suggested data sources. The data are collected using contracted Performance Measurement Systems (vendors) that develop data collection tools based on the measure specifications. The tools are verified and tested by Joint Commission staff to confirm the accuracy and conformance of the data collection tool with the measure specifications. The vendor may not offer the measure set to hospitals until verification has been passed.

S.25. **Data Source or Collection Instrument** *(available at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
No data collection instrument provided

S.26. **Level of Analysis** *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*
Facility, Population : National

S.27. **Care Setting** *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*
Hospital/Acute Care Facility
If other:

S.28. **COMPOSITE Performance Measure** - Additional Specifications *(Use this section as needed for aggregation and weighting rules, or calculation of individual performance measures if not individually endorsed.)*
Not Applicable

2a. Reliability – See attached Measure Testing Submission Form
2b. Validity – See attached Measure Testing Submission Form
PC-05_0480_MeasureTesting_MSF5.0_Data-635787043122090020-635827511042903882.doc

NQF #: 0480 **NQF Project:** Perinatal and Reproductive Health Project

2. RELIABILITY & VALIDITY - SCIENTIFIC ACCEPTABILITY OF MEASURE PROPERTIES

Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. **(evaluation criteria)**
Measure testing must demonstrate adequate reliability and validity in order to be recommended for endorsement. Testing may be conducted for data elements and/or the computed measure score. Testing information and results should be entered in the appropriate field. Supplemental materials may be referenced or attached in item 2.1. See guidance on measure testing.

2a2. Reliability Testing. (*Reliability testing was conducted with appropriate method, scope, and adequate demonstration of reliability.*)

2a2.1 Data/Sample *(Description of the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)*:
The PC measure set has been in national use since the 2nd quarter of 2010. It is a requirement of participation in the ORYX initiative that data on all measures in the set are collected. (ORYX is the term used by The Joint Commission to describe the component of the hospital accreditation program which requires data collection and reporting on standardized national performance measures.)
Demographics of organizations collecting and reporting data on these measures are as follows:
163 health care organizations representing various types, locations and sizes:
10 For Profit, 91 Not for Profit, 46 Military Facilities, 9 County, 2 State, 5 Other
15 >=500 beds; 29 250-499 beds; 50 100-249 beds; 69 <100 beds
Located in: AE, AK, AL, AP, AR, AZ, CA, DO, DC, FL, GA, HI, ID, IL, IN, KS, KY, LA, MA, MD, MI, MN, MO, MS, MT, NC, NE, NV, NY, OH,OK, PA, PR, RI, SC, TN, TX, VA, WA, WI, WV
26 performance measurement systems

2a2.2 Analytic Method *(Describe method of reliability testing & rationale)*:
This measure was adapted from NQF-endorsed measure 0480 Exclusive Breastfeeding During Birth Hospitalization. As such, reliability was addressed during the original endorsement. The Joint Commission will be conducting further reliability studies on this measure as well as the entire PC measure set beginning in October 2011.
Currently, these hospitals are supported in their data collection and reporting efforts by 26 contracted performance measurement system (PMS) vendors. It is a contractual requirement of Joint Commission listed vendors that the quality and reliability of data submitted to them by contracted health care organizations must be monitored on a quarterly basis. In addition, The Joint Commission analyzes these data by running 17 quality tests on the data submitted into ORYX. (ORYX is the term used by The Joint Commission to describe the component of the hospital accreditation program which requires data collection and reporting on standardized national performance measures). The following is a list of the major tests done on the submitted ORYX data, taken from the 2011 ORYX Performance Measurement System Requirements manual.

- Transmission of complete data
- Usage of individual core measure data received: To understand if the HCO provides the relevant service to treat the relevant population
- Investigation of aberrant data points
- Verification of patient population and sample size
- Identification of missing data elements
- Validation of the accuracy of target outliers
- Data integrity
- Data corrections

Data Element Agreement Rate:
Inter-rater reliability testing methodology utilized by contracted performance measure system vendors as outlined in the contract is as

follows:

- All clinical data elements and all editable demographic elements are scored.
- All measure data are reabstracted with originally abstracted data having been blinded so that the reabstraction is not biased.
- Reabstracted data are compared with originally abstracted data on a data element by data element basis. A data element agreement rate is calculated. Clinical and demographic data are scored separately, and an overall agreement rate is computed.

2a2.3 Testing Results *(Reliability statistics, assessment of adequacy in the context of norms for the test conducted)*:
Validity (Measure evaluation criterion 2b)
Data element agreement rates were reported to The Joint Commission for 1Q11. This reflects the findings of106 hospitals, comprising 26,302 records (100% sample). The following table delineates calculated agreement rates for individual data elements that are used to compute measure rates for PC-05.

Data Elements with a Mismatch - Newborn	total n	total d	rate
Admission Date	661	662	99.85%
Admission to NICU	571	576	99.13%
Admission Type	661	662	99.85%
Exclusive Breast Milk Feeding	513	526	97.53%
Point of Origin for Admission or Visit	671	672	99.85%
Reason for Not Exclusively Feeding Breast Milk	334	342	97.66%

These agreement rates are considered to be well within acceptable levels.

2b. VALIDITY. Validity, Testing, including all Threats to Validity: H☐ M☐ L☐ I ☐

2b1.1 Describe how the measure specifications *(measure focus, target population, and exclusions)* **are consistent with the evidence cited in support of the measure focus (***criterion 1c)* **and identify any differences from the evidence:**
This measure focuses on the rate of newborns that were exclusively fed breast milk during the entire birth hospitalization. The literature supports the focus on healthy term newborns that were exclusively fed breast milk. This measure excludes patients diagnosed with galactosemia, receiving TPN, experienced death or requiring a higher level of care due to illness or prematurity resulting in an admission to the NICU or transfer to another hospital, since these are contraindications supported in the literature. Also excluded from the measure are patients with a length of stay greater than 120 days, and those enrolled in a clinical trial. These exclusions are not addressed in the literature, but are included for this measure in order to harmonize with other CMS/Joint Commission aligned measures.

2b2. Validity Testing. (*Validity testing was conducted with appropriate method, scope, and adequate demonstration of validity.)*

2b2.1 Data/Sample *(Description of the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)*:
The PC measure set has been in national use since the 2nd quarter of 2010. Demographics of organizations collecting and reporting data on these measures are as follows:
163 health care organizations representing various types, locations and sizes:
10 For Profit, 91 Not for Profit, 46 Military Facilities, 9 County, 2 State, 5 Other
15 >=500 beds; 29 250-499 beds; 50 100-249 beds; 69 <100 beds

Located in: AE, AK, AL, AP, AR, AZ, CA, DO, DC, FL, GA, HI, ID, IL, IN, KS, KY, LA, MA, MD, MI, MN, MO, MS, MT, NC, NE, NV, NY, OH,OK, PA, PR, RI, SC, TN, TX, VA, WA, WI, WV
26 performance measurement systems

2b2.2 Analytic Method *(Describe method of validity testing and rationale; if face validity, describe systematic assessment)*:
Since the measure has been in national use, continued face validity of the measure has been determined through analysis of feedback from measure users. The Joint Commission provides a web-based application with which measure users can provide feedback regarding appropriateness of measure specifications, request clarification of specifications, and/or provide other comments pertinent to the measure. This feedback is systematically continually reviewed in order to identify trends and to identify areas of the measure specifications that require clarification or revision. Additionally, Joint Commission staff continually monitors the national literature and environment in order to assess continued validity of this measure.

In addition, The Joint Commission will begin reliability site visits this year. During the site visits, Joint Commission staff will conduct focus group interviews with hospital staff working with the PC measures to obtain feedback regarding the validity of the measures and

suggestions for further refinement of the specifications.

2b2.3 **Testing Results** *(Statistical results, assessment of adequacy in the context of norms for the test conducted; if face validity, describe results of systematic assessment)***:**
Analysis of feedback obtained via our automated feedback system reveals slightly more than 130 submissions regarding specifications for this measure since its implementation in 2010. Predominant themes of these submissions involved questions regarding clarification of the data elements Exclusive Breast Milk Feeding regarding the definition and Reason for Not exclusively Feeding Breast Milk as to why additional newborn conditions were not considered exclusions. Also the data element Discharge from NICU was changed to Admission to NICU based on feedback that some hospitals sent newborns to a step-down unit from the NICU prior to discharge. Additional notes for abstractors were added to the data elements for clarification. Other notes for abstractors were added to the data element admission date to clarify the date of delivery is used as the admission date and not the date of the order written to admit. The denominator statement and algorithm were changed to single term newborns discharged from the hospital to capture healthy newborns. In addition, the denominator excluded population and algorithm were revised to capture premature newborns with an additional ICD-9-CM diagnosis code table and newborns transferred to another hospital.

POTENTIAL THREATS TO VALIDITY. (*All potential threats to validity were appropriately tested with adequate results.*)

2b3. Measure Exclusions. (*Exclusions were supported by the clinical evidence in 1c or appropriately tested with results demonstrating the need to specify them.*)

2b3.1 **Data/Sample for analysis of exclusions** *(Description of the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)***:**
The PC measure set has been in national use since the 2nd quarter of 2010. Demographics of organizations collecting and reporting data on these measures are as follows:
163 health care organizations representing various types, locations and sizes:
10 For Profit, 91 Not for Profit, 46 Military Facilities, 9 County, 2 State, 5 Other
15 >=500 beds; 29 250-499 beds; 50 100-249 beds; 69 <100 beds
Located in: AE, AK, AL, AP, AR, AZ, CA, DO, DC, FL, GA, HI, ID, IL, IN, KS, KY, LA, MA, MD, MI, MN, MO, MS, MT, NC, NE, NV, NY, OH,OK, PA, PR, RI, SC, TN, TX, VA, WA, WI, WV
26 performance measurement systems

2b3.2 **Analytic Method** (*Describe type of analysis and rationale for examining exclusions, including exclusion related to patient preference)***:**
Measure exclusions that were not derived directly from the evidence are presented below. Please note that these are population exclusions that are necessary to ensure consistency in all measures in this 5 measure set.

These exclusions were analyzed for frequency of occurrence. An issue that is of great concern to users of this measure is that due to the presence of exceptions to the measure, attainment of a 100% measure rate is not possible. Because of the role of this measure in the current Joint Commission accreditation process, this is especially troubling to measure users. This concern is the basis for a number of the non-evidence-based exclusions to these measures. Additional reasons for these population exclusions are enumerated in our response to section 2b1.1 above. The following measure exclusions that were not derived directly from the evidence are as follows:
1. Patients with LOS <120 days
2. Patients enrolled in clinical trials

2b3.3 **Results** *(Provide statistical results for analysis of exclusions, e.g., frequency, variability, sensitivity analyses):*
N=353,671
1. Patients who have a length of stay (LOS) greater than 120 days =0%
2. Patients enrolled in clinical trials =0%

2b4. Risk Adjustment Strategy. (*For outcome measures, adjustment for differences in case mix (severity) across measured entities was appropriately tested with adequate results.*)

2b4.1 **Data/Sample** *(Description of the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)***:**
Not Applicable

2b4.2 **Analytic Method (***Describe methods and rationale for development and testing of risk model or risk stratification including*

selection of factors/variables):
Not Applicable

2b4.3 **Testing Results** *(Statistical risk model: Provide quantitative assessment of relative contribution of model risk factors; risk model performance metrics including cross-validation discrimination and calibration statistics, calibration curve and risk decile plot, and assessment of adequacy in the context of norms for risk models. Risk stratification: Provide quantitative assessment of relationship of risk factors to the outcome and differences in outcomes among the strata)*:
Not Applicable

2b4.4 **If outcome or resource use measure is not risk adjusted, provide rationale and analyses to justify lack of adjustment:**
Not Applicable

2b5. Identification of Meaningful Differences in Performance. (*The performance measure scores were appropriately analyzed and discriminated meaningful differences in quality.*)

2b5.1 **Data/Sample** *(Describe the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)*:
As previously noted the PC measure set has been in national use since the 2nd quarter of 2010. Demographics of organizations collecting and reporting data on these measures are as follows:
163 health care organizations representing various types, locations and sizes:
10 For Profit, 91 Not for Profit, 46 Military Facilities, 9 County, 2 State, 5 Other
15 >=500 beds; 29 250-499 beds; 50 100-249 beds; 69 <100 beds
Located in: AE, AK, AL, AP, AR, AZ, CA, DO, DC, FL, GA, HI, ID, IL, IN, KS, KY, LA, MA, MD, MI, MN, MO, MS, MT, NC, NE, NV, NY, OH,OK, PA, PR, RI, SC, TN, TX, VA, WA, WI, WV
26 performance measurement systems

2b5.2 **Analytic Method** *(Describe methods and rationale to identify statistically significant and practically/meaningfully differences in performance)*:
The method used to analyze meaningful differences in performance at The Joint Commission is Target Analysis. The object of target analysis is to compare a health care organizations (HCO) data against a comparative norm for the purpose of evaluating performance improvement opportunities. When an organization's performance level is statistically significantly different from a comparative norm, it is considered a statistical deviation. A statistical deviation may be desirable or undesirable depending on the "direction of improvement" of the measure.
There are two components to the target analysis methodology used at The Joint Commission. Given the national average for a performance measure, a target range is constructed. Using generalized linear mixed models methodology (also known as hierarchical models), a predicted estimate of an HCO's performance, with a corresponding 95% confidence interval, is generated. This confidence interval is compared to the target range, to determine the HCOs' rating. The estimate of the organization's true performance is based on both the data from that organization and on data from the entire set of reporting organizations.

2b5.3 **Results** *(Provide measure performance results/scores, e.g., distribution by quartile, mean, median, SD, etc.; identification of statistically significant and meaningfully differences in performance)*:
PC-05 Distribution of Outliers

2011 1st Quarter Data:
Scores on this measure: N=161, Mean 48.33%, SD 0.23493
10th Percentile= 19.23%
25th Percentile= 31.88%
50th Percentile= 50%
75th Percentile= 63.6%
90th Percentile= 78.95%

4 (2.48%) Favorable – results statistically significantly higher than the national rate
119 (73.91%) Neutral – results not significantly different from target range
38 (23.6%) Undesirable –results statistically significantly lower than the national rate

2b6. Comparability of Multiple Data Sources/Methods. (*If specified for more than one data source, the various approaches result in comparable scores.*)

2b6.1 Data/Sample *(Describe the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)*:
Multiple data sources are not used for this measure.

2b6.2 Analytic Method *(Describe methods and rationale for testing comparability of scores produced by the different data sources specified in the measure)*:
Not Applicable

2b6.3 Testing Results *(Provide statistical results, e.g., correlation statistics, comparison of rankings; assessment of adequacy in the context of norms for the test conducted)*:
Not Applicable

2c. Disparities in Care: H☐ M☐ L☐ I ☐ NA☐ (*If applicable, the measure specifications allow identification of disparities.*)

2c.1 If measure is stratified for disparities, provide stratified results *(Scores by stratified categories/cohorts)*: The measure is not stratified for disparities.

2c.2 If disparities have been reported/identified (e.g., in 1b), but measure is not specified to detect disparities, please explain:
Although rates of exclusive breastfeeding were significantly lower among black infants (compared with white infants) and infants born to unmarried mothers (compared with married mothers, this measure is not stratified for these elements. The Joint Commission does not currently capture date elements for race or ethnicity because these data elements have not been shown to be reliably collectable due to the fact that no national standardized definitions exist for these data elements. Also, not all hospitals collect race and ethnicity. In the future, it may be feasible for The Joint Commission to explore how race and ethnicity and other relevant disparity data might be collected reliably in the future.

2.1-2.3 Supplemental Testing Methodology Information:

Steering Committee: Overall, was the criterion, *Scientific Acceptability of Measure Properties*, met?
(*Reliability and Validity must be rated moderate or high*) Yes☐ No☐
Provide rationale based on specific subcriteria:

If the Committee votes No, STOP

NEW (since 2012 endorsement): Data for Empirical Testing

2b2. VALIDITY TESTING
2b2.1. What level of validity testing was conducted? (*may be one or both levels*)

- ☐ **Critical data elements** (*data element validity must address ALL critical data elements*)
- ☐ **Performance measure score**
- ☒ **Empirical validity testing**
- ☐ **Systematic assessment of face validity of performance measure score as an indicator** of quality or resource use (*i.e., is an accurate reflection of performance on quality or resource use and can distinguish good from poor performance*)

2b2.2. For each level of testing checked above, describe the method of validity testing and what it tests (*describe the steps—do not just name a method; what was tested, e.g., accuracy of data elements compared to authoritative source, relationship to another measure as expected; what statistical analysis was used)*

ICD-9 to ICD-10 Conversion Process:
The goal was to convert ICD-9 to ICD-10 equivalent codes, consistent with the clinical intent of the original measure specifications. The Joint Commission worked with a certified coding expert throughout the conversion process. The 3M Coding Conversion Tool was utilized, including forward mapping of ICD-9 codes to ICD-10 codes as well as reverse mapping from ICD-10 to ICD-9 to ensure appropriateness. MSDRGs and instructions

in the tabular index were also examined to ensure appropriate code mapping. Crosswalks comprising ICD-9 codes mapped to ICD-10 codes were created and reviewed by members of the Technical Advisory Panel, CMS subcontractors, and performance measurement system vendors prior to being posted for a 12 month public comment period. Feedback from the field indicated that the crosswalks generally were mapped correctly. Minor modifications to the code tables were made as needed. Final code tables were published in early 2015, well in advance of the mandated date of October 1, 2015.

Newborns with Exclusive Breast Feeding - Patient Age at admission (*Admission Date – Birthdate*) ≤ 2 days, Length of Stay (*Discharge Date - Admission Date*) ≤ 120 days, an *ICD-9-CM Principal* as defined in Appendix A, Table 11.20.1, **NO** *ICD-9-CM Other Diagnosis Codes* as defined in Appendix A, Table 11.21, **NO** *ICD-9-CM-Principal* or *Other Procedure Code* as defined in Appendix A, Table 11.22 and **NO** *ICD-9-CM Other Diagnosis Codes* as defined in Appendix A, Table 11.23 are included in this subpopulation and are eligible to be sampled.
PC-05 measure belongs to this sub population

The data used to measure the validity of the PC measure are comprised of data from the third and fourth quarters of 2014, and the first and second quarters of 2015. 1,345 hospitals submitted 2,695,467 inpatient records for all the elected PC measures. The hospitals included in the analysis reported one year of data and had 30 or more denominator cases in the analysis period.
Measure convergent validity for PC-05 was assessed using hospitals patient level data from The Joint Commission warehouse. Measure specifications, including population identification, numerator and denominator statements, exclusions, and data elements and their definitions were found to be understandable, retrievable, and relevant in previous validity testing.

2b2.3. What were the statistical results from validity testing? (*e.g., correlation; t-test*)
Descriptive statistics for sub population 3 - PC-05 Newborns with Breast Feeding

N=1,352 hospitals
n= 775,909 records submitted

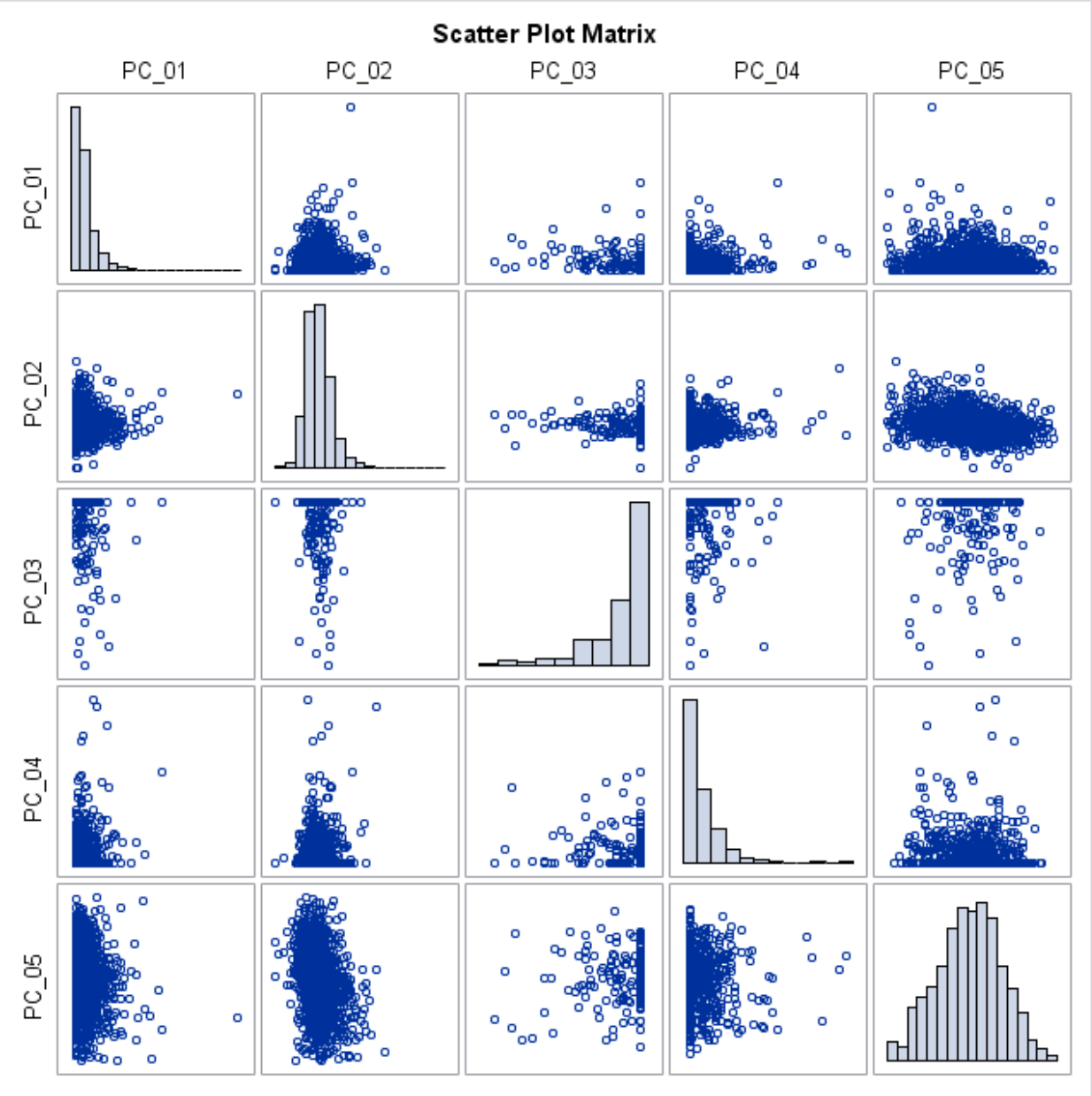
Analysis of hospitals' rate in conjunction with PC-05 measure

PC-05
Percentile 10%: 23%
Percentile 25%: 37%
Median: 50%
Percentile 75%: 63%
Percentile 90%: 75%

Simple Statistics

Variable	N	Mean	Std Dev	Median	Minimum	Maximum
PC_01	1237	0.02753	0.03803	0.01734	0	0.51240
PC_02	1345	0.26287	0.07974	0.25410	0	1.00000
PC_03	162	0.97762	0.03311	0.99425	0.84615	1.00000
PC_04	523	0.05267	0.08432	0.02203	0	0.66129
PC_05	1352	0.49198	0.19284	0.50190	0.00317	1.00000

Spearman Correlation Coefficients Prob > |r| under H0: Rho=0 Number of Observations					
	PC_01	PC_02	PC_03	PC_04	PC_05
PC_01	1.00000 1237	0.06843 0.0163 1231	-0.26960 0.0006 159	0.10724 0.0169 496	-0.03538 0.2137 1237
PC_02	0.06843 0.0163 1231	1.00000 1345	-0.18318 0.0196 162	0.02807 0.5218 523	-0.32009 <.0001 1343
PC_03	-0.26960 0.0006 159	-0.18318 0.0196 162	1.00000 162	-0.03117 0.7030 152	0.07729 0.3283 162
PC_04	0.10724 0.0169 496	0.02807 0.5218 523	-0.03117 0.7030 152	1.00000 523	-0.03560 0.4165 523
PC_05	-0.03538 0.2137 1237	-0.32009 <.0001 1343	0.07729 0.3283 162	-0.03560 0.4165 523	1.00000 1352



The Spearman rank-order correlation is a nonparametric measure of association based on the ranks of the data values by measure PC-05 and hospitals. We used this methodology because of the skewness of the distribution of the measure rates.

2b2.4. What is your interpretation of the results in terms of demonstrating validity? (i.*e., what do the results mean and what are the norms for the test conducted?*)

The correlation of PC-05 with PC-02 is moderate and statistically significant. The other correlations with the other PC measures are relatively weak and not significant. Performance of hospitals on this measure varied widely, 90% of the hospital measure rates fall between 23 and 75%, indicating that there is much room for improvement on this measure.

2b3. EXCLUSIONS ANALYSIS .
NA ☐ no exclusions — *skip to section 2b4*

There were 775,909 admissions selected from the initial cohort. From among the 775,909 admissions in 1,352 hospitals, the descriptive statistics are given below.

The following exclusions were analyzed by subpopulation and measure for frequency and variability across providers:

Excluded Populations:

- Admitted to the Neonatal Intensive Care Unit (NICU) at this hospital during the hospitalization
- *ICD-9-CM Other Diagnosis Codes* for galactosemia as defined in Appendix A, Table 11.21
- *ICD-9-CM Principal Procedure Code or ICD-9-CM Other Procedure Codes* for parenteral infusion as defined in Appendix A, Table 11.22
- Experienced death
- Length of Stay >120 days
- Enrolled in clinical trials
- Documented *Reason for Not Exclusively Feeding Breast Milk*
- Patients transferred to another hospital
- *ICD-9-CM Other Diagnosis Codes* for premature newborns as defined in Appendix A Table 11.23

2b3.2. What were the statistical results from testing exclusions? (*include overall number and percentage of individuals excluded, frequency distribution of exclusions across measured entities, and impact on performance measure scores*)

There were 775,909 admissions selected from the initial cohort. From among the 775,909 admissions in 1,352 hospitals, the descriptive statistics are given below.

Exclusion Subpopulation 3 - PC-05
Exclusion: Admitted to the Neonatal Intensive Care Unit (NICU) at this hospital during the hospitalization
Overall Number of Occurrences n = 40,754
Overall Occurrence Percentage: 5.25%
Minimum: 0 %
10th Percentile: 0%
Median: 4.25%
90th Percentile: 11.2%
Maximum: 69%

Exclusion: *ICD-9-CM Other Diagnosis Codes* for galactosemia as defined in Appendix A, Table 11.21
No observations noted

Exclusion: *ICD-9-CM Principal Procedure Code or ICD-9-CM Other Procedure Codes* for parenteral infusion as defined in Appendix A, Table 11.22
No observations noted

Exclusion: *Patients who expire during the hospital stay*
Overall Number of Occurrences n = 404
Overall Occurrence Percentage: 0.05%
Minimum: 0%
10th Percentile: 0%
Median: 0%
90th Percentile: 0.2%
Maximum: 1.9%

Exclusion: *Length* of Stay >120 days
No observations noted

Exclusion: *Patients enrolled in clinical trials*
Overall Number of Occurrences n = 248
Overall Occurrence Percentage: .03%
Minimum 0%
10th Percentile: 0%
Median: 0%
90th Percentile: 0%
Maximum: 32%

Exclusion: *Documented Reason for Not Exclusively Feeding Breast Milk*
Overall Number of Occurrences n = 7,282
Overall Occurrence Percentage: 0 .94%
Minimum 0%
10th Percentile: 0%
Median: 0.6%
90th Percentile: 2.3%
Maximum: 17.2%

Exclusion: *Patients transferred to another hospital;*
Overall Number of Occurrences n = 459
Overall Occurrence Percentage: 0.06%
Minimum 0%
10th Percentile: 0%
Median: 0%
90th Percentile: 0.2%
Maximum: 5.5%

Exclusion: *ICD-9-CM Other Diagnosis Codes for premature newborns as defined in Appendix A, Table 11.23*
No observations noted

2b3.3. What is your interpretation of the results in terms of demonstrating that exclusions are needed to prevent unfair distortion of performance results? (*i.e., the value outweighs the burden of increased data collection and analysis. Note: **If patient preference is an exclusion**, the measure must be specified so that the effect on the performance score is transparent, e.g., scores with and without exclusion*)

The overall frequency of exclusions is low for those in the measure denominator. The difference between the 10th and 90th percentiles of the distribution of exclusion rates is narrow indicating that the occurrence is random and likely would not bias performance results.

It is believed that all of the exclusions should be retained for the following reasons:

Exclusion: Admitted to the Neonatal Intensive Care Unit (NICU) at this hospital during the hospitalization
Rationale: Newborns admitted to the NICU are excluded from the measure, since PC-05 only includes healthy term newborns.

Exclusion: *ICD-9-CM Other Diagnosis Codes* for galactosemia as defined in Appendix A, Table 11.21
Rationale: Newborns diagnosed with galactosemia are excluded from the measure, since breast milk is contraindicated due to the newborn's inability to digest milk proteins.

Exclusion: *ICD-9-CM Principal Procedure Code or ICD-9-CM Other Procedure Codes* for parenteral infusion as defined in Appendix A, Table 11.22
Rationale: Newborns receiving parenteral infusions are excluded from the measure, since these newborns are NPO and no oral feedings would be given.

Exclusion: Patients who expired

Rationale: Patients who expire are not eligible to be in this measure

Exclusion: Length of Stay >120 days
Rationale: Included for this measure in order to harmonize with other CMS/Joint Commission aligned measures.

Exclusion: Patients enrolled in a Clinical Trial
Rationale: Newborns enrolled in clinical trials are excluded from the measure, since a clinical trial involves an intervention and control group which may impact the ability to provide breast milk feedings.

Exclusion: *Documented Reason for Not Exclusively Feeding Breast Milk*
Rationale: This data element has been removed effective with 10/1/15 discharges. Exclude cases when there were certain maternal medical reasons that the newborn could not be fed breast milk.

Exclusion: Patients transferred to another hospital
Rationale: : Newborns transferred to another hospital are excluded from the measure, since most of these newborns are NPO and are being transferred to a higher level of care due to medical conditions

Exclusion: *ICD-9-CM Other Diagnosis Codes* for premature newborns as defined in Appendix A Table 11.23
Rationale: Newborns with prematurity codes are excluded from the measure, since these are not term newborns which is the population of interest. This was replaced with the new data element Term Newborn effective with 10/1/15 discharges in order to exclude pre-term newborns by gestational age <37 weeks and/or descriptors for pre-term.

3. Feasibility

Extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

3a. Byproduct of Care Processes

For clinical measures, the required data elements are routinely generated and used during care delivery (e.g., blood pressure, lab test, diagnosis, medication order).

3a.1. **Data Elements Generated as Byproduct of Care Processes.**

generated by and used by healthcare personnel during the provision of care, e.g., blood pressure, lab value, medical condition, Coded by someone other than person obtaining original information (e.g., DRG, ICD-9 codes on claims), Abstracted from a record by someone other than person obtaining original information (e.g., chart abstraction for quality measure or registry)

If other:

3b. Electronic Sources

The required data elements are available in electronic health records or other electronic sources. If the required data are not in electronic health records or existing electronic sources, a credible, near-term path to electronic collection is specified.

3b.1. **To what extent are the specified data elements available electronically in defined fields?** *(i.e., data elements that are needed to compute the performance measure score are in defined, computer-readable fields)*

Some data elements are in defined fields in electronic sources

3b.2. **If ALL the data elements needed to compute the performance measure score are not from electronic sources, specify a credible, near-term path to electronic capture, OR provide a rationale for using other than electronic sources.**

The Joint Commission recognizes that not all hospitals currently have the capacity to abstract the electronic version of this measure, so continues to offer this chart abstracted version which allows for data capture from unstructured data fields. All data elements needed to compute the PC-05 performance measure score have been retooled for capture from electronic sources. Annual updates are performed to match the eCQM specifications to the current version of the chart-abstracted specifications.

3b.3. **If this is an eMeasure, provide a summary of the feasibility assessment in an attached file or make available at a measure-specific URL.**

Attachment:

3c. Data Collection Strategy

Demonstration that the data collection strategy (e.g., source, timing, frequency, sampling, patient confidentiality, costs associated with fees/licensing of proprietary measures) can be implemented (e.g., already in operational use, or testing demonstrates that it is ready to put into operational use). For eMeasures, a feasibility assessment addresses the data elements and measure logic and demonstrates the eMeasure can be implemented or feasibility concerns can be adequately addressed.

3c.1. **Describe what you have learned/modified as a result of testing and/or operational use of the measure regarding data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, time and cost of data collection, other feasibility/implementation issues.**

IF a PRO-PM, consider implications for both individuals providing PROM data (patients, service recipients, respondents) and those whose performance is being measured.

At the present time, hospitals using this performance measure generally collect measure data via manual review of the paper medical record, the EMR or a combination of both. Collected data are submitted to The Joint Commission on a quarterly basis, by way of contracted performance measurement system vendors, as described previously. Specifications for this measure are freely available to anyone who wishes to use the measure. Feedback from hospitals using this measure indicates that required data elements are generally available in the medical record, and measure specifications are robust and easy to understand. As described above, as feedback from measure users has indicated the need for clarification or revision of measure specifications, this has taken place.

3c.2. **Describe any fees, licensing, or other requirements to use any aspect of the measure as specified** *(e.g., value/code set, risk model, programming code, algorithm)*.

There are no fees or licensing requirements to use the Joint Commission performance measures, all of which are in the public domain.

Usability and Use

4. Usability and Use

Extent to which potential audiences (e.g., consumers, purchasers, providers, policy makers) are using or could use performance results for both accountability and performance improvement to achieve the goal of high-quality, efficient healthcare for individuals or populations.

4a. Accountability and Transparency

Performance results are used in at least one accountability application within three years after initial endorsement and are publicly reported within six years after initial endorsement (or the data on performance results are available). If not in use at the time of initial endorsement, then a credible plan for implementation within the specified timeframes is provided.

4.1. Current and Planned Use

NQF-endorsed measures are expected to be used in at least one accountability application within 3 years and publicly reported within 6 years of initial endorsement in addition to performance improvement.

Planned	Current Use (for current use provide URL)
	Public Reporting Quality Check® Regulatory and Accreditation Programs Hospital Accreditation Program http://jointcommission.org Quality Improvement with Benchmarking (external benchmarking to multiple organizations) America's Hospitals: Improving Quality and Safety – The Joint Commission's Annual Report http://www.jointcommission.org/annualreport.aspx Quality Improvement (Internal to the specific organization) Perinatal Care Certification http://www.jointcommission.org/certification/perinatal_care_certification.aspx

4a.1. For each CURRENT use, checked above, provide:

- Name of program and sponsor
- Purpose
- Geographic area and number and percentage of accountable entities and patients included

- Name of program and sponsor: Quality Check®; The Joint Commission
- Purpose: A public website that allows consumers to: search for accredited and certified organizations by city and state, by name or by zip code (up to 250 miles); find organizations by type of service provided within a geographic area; download free hospital performance measure results; and, print a list of Joint Commission certified disease-specific care programs and health care staffing firms.
- Geographic area and number and percentage of accountable entities and patients included: Nationwide; 3300 Joint Commission-accredited hospitals (2014)

- Name of program and sponsor Hospital Accreditation Program; The Joint Commission
- Purpose: An accreditation program that recognizes hospitals that meet standard requirements to provide safe and effective patient care.
- Geographic area and number and percentage of accountable entities and patients included Nationwide; 3300 Joint Commission-accredited hospitals (2014)

- Name of program and sponsor America's Hospitals: Improving Quality and Safety – The Joint Commission's Annual Report ; The Joint Commission

- Purpose: The annual report summarizes the performance of Joint Commission-accredited hospitals on 46 accountability measures of evidence-based care processes closely linked to positive patient outcomes, and provides benchmarks from Top Performer on Key Quality Measures® hospitals.
- Geographic area and number and percentage of accountable entities and patients included Nationwide; 3300 Joint Commission-accredited hospitals (2014)

- Name of program and sponsor Perinatal Care Certification; The Joint Commission
- Purpose: A certification program that recognizes hospitals that have achieved integrated, coordinated, patient-centered care for clinically uncomplicated pregnancies and births.
- Geographic area and number and percentage of accountable entities and patients included Nationwide; Twelve Joint Commission-accredited hospitals (2016)

4a.2. **If not currently publicly reported OR used in at least one other accountability application (e.g., payment program, certification, licensing) what are the reasons?** (*e.g., Do policies or actions of the developer/steward or accountable entities restrict access to performance results or impede implementation?*)
Not Applicable

4a.3. **If not currently publicly reported OR used in at least one other accountability application, provide a credible plan for implementation within the expected timeframes -- any accountability application within 3 years and publicly reported within 6 years of initial endorsement.** (*Credible plan includes the specific program, purpose, intended audience, and timeline for implementing the measure within the specified timeframes. A plan for accountability applications addresses mechanisms for data aggregation and reporting.*)
Not Applicable

4b. Improvement

Progress toward achieving the goal of high-quality, efficient healthcare for individuals or populations is demonstrated. If not in use for performance improvement at the time of initial endorsement, then a credible rationale describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.

4b.1. **Progress on Improvement. (Not required for initial endorsement unless available.)**
Performance results on this measure (current and over time) should be provided in 1b.2 and 1b.4. Discuss:

- **Progress (trends in performance results, number and percentage of people receiving high-quality healthcare)**
- **Geographic area and number and percentage of accountable entities and patients included**

The rate of exclusive breast milk feeding slowly improved from 40.9% in 2010 with 165 hospitals reporting to 49.4% in 2014 with 1386 hospitals reporting based on Joint Commission ORYX performance measurement data. Beginning with January 1, 2016 discharges, an additional 821 accredited hospitals will begin reporting the data. The new reporting requirement will capture approximately 80% of the accredited hospitals with maternity services in the US. The most improvement in 2014 was noted for the 90th percentile (74.3%) hospitals. This underscores the importance to continue to monitor progress towards improving the rate in order to reach the performance goal of 70%.

4b.2. **If no improvement was demonstrated, what are the reasons? If not in use for performance improvement at the time of initial endorsement, provide a credible rationale that describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.**
Not Applicable

4c. Unintended Consequences

The benefits of the performance measure in facilitating progress toward achieving high-quality, efficient healthcare for individuals or populations outweigh evidence of unintended negative consequences to individuals or populations (if such evidence exists).

4c.1. **Were any unintended negative consequences to individuals or populations identified during testing; OR has evidence of unintended negative consequences to individuals or populations been reported since implementation? If so, identify the negative unintended consequences and describe how benefits outweigh them or actions taken to mitigate them.**
Unintended Consequence: Data abstraction for the mother's initial feeding plan for PC-05a: Exclusive Breast Milk Feeding Considering Mother's Initial Feeding Plan did not follow normal workflow patterns and greatly increased the burden of data abstraction for hospitals. Numerous stakeholders also expressed concern that undecided mothers were often choosing both formula and breast milk, and it was perceived that these mothers were not receiving the same level of support as mothers who had chosen to feed breast milk only.

Mitigating Action: Retired the sub-measure PC-05a: Exclusive Breast Milk Feeding Considering Mother's Initial Feeding Plan.

Unintended Consequence: Late preterm newborns were not being routinely excluded from the denominator population via diagnosis codes for premature newborns on Table 11.23.
Mitigating Action: Table 11.23 was removed and a new denominator data element Term Newborn was added, so that now only newborns who were term or =>37 weeks gestation completed were included.

Unintended Consequence: Reviewing medical records for maternal medical conditions as reasons for not exclusively feeding breast milk was greatly increasing the burden of data abstraction based on feedback from hospitals.
Mitigating Action: The denominator data element Reason For Not Exclusively Feeding Breast Milk was removed from the denominator excluded population in order to simplify data abstraction.

5. Comparison to Related or Competing Measures

If a measure meets the above criteria and there are endorsed or new related measures (either the same measure focus or the same target population) or competing measures (both the same measure focus and the same target population), the measures are compared to address harmonization and/or selection of the best measure.

5. Relation to Other NQF-endorsed Measures
Are there related measures (conceptually, either same measure focus or target population) or competing measures (conceptually both the same measure focus and same target population)? If yes, list the NQF # and title of all related and/or competing measures.
No

5.1a. **List of related or competing measures (selected from NQF-endorsed measures)**

5.1b. **If related or competing measures are not NQF endorsed please indicate measure title and steward.**

5a. Harmonization

The measure specifications are harmonized with related measures;
OR
The differences in specifications are justified

5a.1. **If this measure conceptually addresses EITHER the same measure focus OR the same target population as NQF-endorsed measure(s):**
Are the measure specifications completely harmonized?

5a.2. **If the measure specifications are not completely harmonized, identify the differences, rationale, and impact on interpretability and data collection burden.**
Not Applicable

5b. Competing Measures

The measure is superior to competing measures (e.g., is a more valid or efficient way to measure);
OR
Multiple measures are justified.

5b.1. **If this measure conceptually addresses both the same measure focus and the same target population as NQF-endorsed measure(s):**
Describe why this measure is superior to competing measures (e.g., a more valid or efficient way to measure quality); OR provide a rationale for the additive value of endorsing an additional measure. (Provide analyses when possible.)
Not Applicable

Appendix

A.1 Supplemental materials may be provided in an appendix. All supplemental materials (such as data collection instrument or methodology reports) should be organized in one file with a table of contents or bookmarks. If material pertains to a specific submission form number, that should be indicated. Requested information should be provided in the submission form and required attachments. There is no guarantee that supplemental materials will be reviewed.

Available at measure-specific web page URL identified in S.1 **Attachment:**

Contact Information

Co.1 **Measure Steward (Intellectual Property Owner):** The Joint Commission
Co.2 **Point of Contact:** Ann, Watt, awatt@jointcommission.org, 630-792-5944-
Co.3 **Measure Developer if different from Measure Steward:** The Joint Commission

Co.4 **Point of Contact:** Ann, Watt, awatt@jointcommission.org, 630-792-5944-

Additional Information

Ad.1 **Workgroup/Expert Panel involved in measure development**

Provide a list of sponsoring organizations and workgroup/panel members' names and organizations. Describe the members' role in measure development.

Michael Ross, MD, MPH (Chair)
Harbor-UCLA Medical Center
Torrance, CA

Wanda Barfield, MD, MPH
Centers for Disease Control and Prevention
Atlanta, GA

Kenneth E. Brown, MD, MBA, FACOG, FACHE
Woman´s Hospital
Lafayette, LA

Martin McCaffrey, MD
UNC North Carolina Children's Hospital
Chapel Hill, NC

Cathy Collins-Fulea, MSN, CNM
Henry Ford Hospital
Detroit, MI

Janet H. Muri, MBA
National Perinatal Information Center/
Quality Analytic Services
Providence, RI

Kathleen Simpson, PhD, RNC, FAAN
St. John's Mercy Medical Center
St. Louis, MO

Michael Socol, MD
Northwestern University Medical School
Chicago, IL

Rebecca Zimmermann, MPP
America's Health Insurance Plans
Washington, DC

The technical advisory panel (TAP) members determined priority areas that could be evaluated to improve care related to perinatal care during the development timeframe. After implementation, minor revisions, acknowledged by TAP representatives, were made to improve clarity. Hospital feedback will be reviewed during the reliability testing phase of the project to assist the TAP in making the final measure recommendations.

Measure Developer/Steward Updates and Ongoing Maintenance

Ad.2 **Year the measure was first released:** 2010
Ad.3 **Month and Year of most recent revision:** 10, 2015
Ad.4 **What is your frequency for review/update of this measure?** Biannual
Ad.5 **When is the next scheduled review/update for this measure?** 02, 2016

associated documentation based on the published manual production timelines. Ad.7 **Disclaimers:**
Ad.8 **Additional Information/Comments:**

MEASURE WORKSHEET

This document summarizes the evaluation of the measure as it progresses through NQF's Consensus Development Process (CDP). The information submitted by measure developers/stewards is included after the Brief Measure Information, Preliminary Analysis, and Pre-meeting Public and Member Comments sections.

To navigate the links in the worksheet: Ctrl + click link to go to the link; ALT + LEFT ARROW to return

Brief Measure Information

NQF #: 0483
Measure Title: Proportion of infants 22 to 29 weeks gestation screened for retinopathy of prematurity.
Measure Steward: Vermont Oxford Network
Brief Description of Measure: Proportion of infants born from 22 weeks, 0 days to 29 weeks, 6 days gestational age who were in the reporting hospital at the postnatal age recommended for screening for retinopathy of prematurity (ROP) by the American Academy of Pediatrics (AAP) and who received a retinal examination for ROP prior to discharge.
Developer Rationale: The goal of this measure is to increase proportion of high risk infants who receive retinal screening on the recommended schedule.

Numerator Statement: Number of infants born from 22 weeks, 0 days to 29 weeks, 6 days gestational age who were in the reporting hospital at the postnatal age recommended for ROP screening by the AAP and who received a retinal exam for ROP prior to discharge
Denominator Statement: All eligible infants born from 22 weeks, 0 days to 29 weeks, 6 days gestational age who were in the reporting hospital at the postnatal age recommended for ROP screening by the AAP
Denominator Exclusions: 1. Infants outside the gestational age range of 22 to 29 weeks
2. Outborn infants admitted to the reporting hospital more than 28 days after birth
3. Outborn infants who have been home prior to admission
4. Infants who die in the delivery room or initial resuscitation area prior to admission to the neonatal intensive care unit
5. Infants not in the reporting hospital at the postnatal age recommended for ROP screening by the AAP

Measure Type: Process
Data Source: Electronic Clinical Data : Registry
Level of Analysis: Facility

IF Endorsement Maintenance – Original Endorsement Date: Oct 24, 2008 **Most Recent Endorsement Date:** Mar 30, 2012

Maintenance of Endorsement -- Preliminary Analysis

To maintain NQF endorsement endorsed measures are evaluated periodically to ensure that the measures still meets the NQF endorsement criteria ("maintenance"). The emphasis for maintaining endorsement is focused on how effective the measure is for promoting improvements in quality. Endorsed measures should have some experience from the field to inform the evaluation. The emphasis for maintaining endorsement is noted for each criterion.

Criteria 1: Importance to Measure and Report

1a. Evidence
Maintenance measures – less emphasis on evidence unless there is new information or change in evidence since the prior evaluation.

1a. Evidence. The evidence requirements for a *process or intermediate outcome* measure is that it is based on a systematic review (SR) and grading of the body of empirical evidence where the specific focus of the evidence matches what is being measured.

The developer provides the following evidence for this measure:

- **Systematic Review of the evidence specific to this measure?** ☒ **Yes** ☐ **No**
- **Quality, Quantity and Consistency of evidence provided?** ☒ **Yes** ☐ **No**
- **Evidence graded?** ☐ **Yes** ☒ **No**

Evidence Summary from 2012

- The developer provides a summary of the structure-process-outcome relationship.
- Evidence is based on clinical practice guidelines, selected individual studies, and a systematic review.
- The developers state that "As far as we are aware the guidelines for timing of retinal examination have not been tested in a randomized trial. However, clear evidence from randomized trials on the effectiveness of surgical or pharmacological intervention for infants with threshold retinopathy of prematurity (ROP) leaves little doubt that appropriately timed retinal exams are required to identify infants who will benefit from treatment."

Changes to evidence from last review

☒ **The developer attests that there have been no changes in the evidence since the measure was last evaluated.**

☐ **The developer provided updated evidence for this measure:**

Updates:

- The developers provided three new studies published in 2014-2016.
- They also updated the form to reflect changes to the guideline.

Exception to evidence

N/A

Guidance from the Evidence Algorithm

For measures that assess performance on a process, it is based on a systematic review (SR) and grading of the BODY of empirical evidence (Box 3) →Yes→Is a summary of the QQC of the body of evidence from a SR provided? → Yes→ Q/H; Q/H; C/H →Rate as High

Questions for the Committee:

- *Although the guidelines have been updated/new studies have been provided, the underlying evidence presented appears to be the directionally the same/stronger since the last NQF endorsement review. Does the Committee agree and so there is no need for repeat discussion and vote on Evidence?*

Preliminary rating for evidence: ☒ **Pass** ☐ **No Pass**

1b. Gap in Care/Opportunity for Improvement and 1b. disparities
Maintenance measures – increased emphasis on gap and variation

1b. Performance Gap. The performance gap requirements include demonstrating quality problems and opportunity for improvement.

- The developer reports "In 2014, 916 hospitals in the Vermont Oxford Network enrolled 40,638 eligible infants with gestational age between 22 and 29 weeks, and 32,325 (79.5%) were in the reporting hospital at the postnatal age recommended for ROP screening. Of those, 6.2% did not receive retinal exams prior to discharge, with the interquartile range among hospitals of 0% to 10% not receiving recommended screenings."
  - The mean has improved over time from 0.901 in 2006 to 0.918 in 2014.
- From 2000 to 2009, rates of any retinal examination among infants weighing 501 to 1500 g and cared for at Vermont Oxford Network members in North America increased from 66.2% to 74.5%

Disparities

- In 2014, there were no differences by maternal race in rates of eye examination at the recommended age: White non-Hispanic (95%); black non-Hispanic (93%); Hispanic (95%); Asian (94%).

Questions for the Committee:

- *Is there a gap in care that warrants a national performance measure?*
- *From the data provided, it appears the gap in performance may be in the low birthweight babies?*
- *Do you expect continued improvements in performance assessed by this measure?*

Preliminary rating for opportunity for improvement: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 1: Importance to Measure and Report (including 1a, 1b, 1c)

1a. Evidence to Support Measure Focus

Comments:

The measure applies directly and is relevant to the desired outcome.

Yes although no randomized trials have been performed

1b. Performance Gap

Comments:

The measure did show improvement from 2000 to 2009. It also documents no differences based on maternal race.

No gap in race/ethnicity appears to be by gestational age.

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability

2a1. Reliability Specifications

Maintenance measures – no change in emphasis – specifications should be evaluated the same as with new measures

2a1. Specifications requires the measure, as specified, to produce consistent (reliable) and credible (valid) results about the quality of care when implemented.

Data source(s): Electronic clinical data – Registry

Specifications:

- The numerator is Number of infants who received a retinal exam for ROP prior to discharge.
- The denominator is All eligible infants born from 22 weeks, 0 days to 29 weeks, 6 days gestational age who were in the reporting hospital at the postnatal age recommended for ROP screening by the AAP
- There are five exclusions:
  - Infants outside the gestational age range of 22 to 29 weeks
  - Outborn infants admitted to the reporting hospital more than 28 days after birth
  - Outborn infants who have been home prior to admission
  - Infants who die in the delivery room or initial resuscitation area prior to admission to the neonatal intensive care unit
  - Infants not in the reporting hospital at the postnatal age recommended for ROP screening by the AAP
- The measure is not risk adjusted, but it is stratified by gestational age, birth location and birth weight category.
- The calculation algorithm is included.
- There is no sampling.
- An attached spreadsheet contains numerous ICD-9 and ICD-10 codes for newborn gestational age, newborn

birth weight, and sepsis, however, VON does not use ICD-9 or ICD-10 codes in its definitions and noted that the supplied codes have not been reviewed or vetted.

Questions for the Committee:

- *Are all the data elements clearly defined? Are all appropriate codes included?*
- *Is the logic or calculation algorithm clear?*
- *Is it likely this measure can be consistently implemented?*

2a2. Reliability Testing Testing attachment
Maintenance measures – less emphasis if no new testing data provided

2a2. Reliability testing demonstrates if the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise enough to distinguish differences in performance across providers.

For maintenance measures, summarize the reliability testing from the prior review:
Specific reliability testing was not discussed.

Describe any updates to testing see below

SUMMARY OF TESTING
Reliability testing level ☒ **Measure score** ☐ **Data element** ☐ **Both**
Reliability testing performed with the data source and level of analysis indicated for this measure ☒ **Yes** ☐ **No**

Method(s) of reliability testing

- The developer reports on a split-half analysis to measure the internal consistency of the measure. They selected hospitals with at least 10 infants, ran 100 samples of patients within each hospital randomly split to two groups, computed hospital rates, and calculated the correlations.

Results of reliability testing

- Aggregate reliability for 2006 – 2014 is provided. The reliability was 0.72 in 2014. Reliability estimates range from 0 (no correlation) to 1 (perfect correlation). Split-half is an appropriate test of reliability.
- The developers conclude that "The correlation coefficients were lower than we expected. It suggests that the definition may not be applied in the same manner across all infants at all hospitals. The coefficients increased as the number of infants at the hospitals increased."

Guidance from the Reliability Algorithm

Precise specifications (Box 1) → empirical testing (box2) → testing of measure score (Box 4) → appropriate method (Box 5) → moderate confidence scores are reliable (Box 6b) →moderate

Questions for the Committee:

- *Is the test sample adequate to generalize for widespread implementation?*
- *Do the results demonstrate sufficient reliability so that differences in performance can be identified?*

Preliminary rating for reliability: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

2b. Validity
Maintenance measures – less emphasis if no new testing data provided

2b1. Validity: Specifications

2b1. Validity Specifications. This section should determine if the measure specifications are consistent with the evidence.

Specifications consistent with evidence in 1a. ☒ **Yes** ☐ **Somewhat** ☐ **No**
Specification not completely consistent with evidence

Question for the Committee:
- *Are the specifications consistent with the evidence?*

2b2. Validity testing

2b2. Validity Testing should demonstrate the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality.

For maintenance measures, summarize the validity testing from the prior review: Face validity assessment only.

Describe any updates to validity testing

SUMMARY OF TESTING
Validity testing level ☐ **Measure score** ☐ **Data element testing against a gold standard** ☐ **Both**

Method of validity testing of the measure score:
- ☒ **Face validity only**
- ☐ **Empirical validity testing of the measure score**

Validity testing method:
- The measure was developed by board certified neonatologists and reviewed by clinical experts in pediatric ophthalmology.
- Vermont Oxford Network's Database Advisory Committee, consisting of national and international experts in the neonatal community, reviews the measure annually.

Validity testing results:
- Results of a systematic assessment of face validity was not provided.

Questions for the Committee:
- *Is the test sample adequate to generalize for widespread implementation?*
- *Do the results demonstrate sufficient validity so that conclusions about quality can be made?*
- *Do you agree that the score from this measure as specified is an indicator of quality?*

2b3-2b7. Threats to Validity

2b3. Exclusions:
- Infants are excluded if they are transferred early (prior to the appropriate time to receive the exam) or if they had the exam, but at the wrong time. The developer reports the following analysis of exclusions:
- The correlation coefficient between the proportion of early transfers and the proportion who received a retinal exam at the recommended post menstrual age was -0.124 (P<0.0001). However, the average proportion of early transfers was 0.72%; over half of the hospitals had an early transfer rate of 0%. Therefore it is difficult to judge the effect of early transfer on eye exams by the size of the correlation coefficient. Among hospitals that had any early transfers, the correlation coefficient was -0.186 (P<0.0001).
- The correlation coefficient between the proportion of infants who had a retinal examination and the proportion who had one at the recommended post menstrual age was 0.833 (P<0.0001). The difference between the proportion of infants who had a retinal examination and those who had one at the recommended post menstrual age was 12% on average, ranging from 0% (all infants had eye exams at the recommended PMA) to 89% who did not have the eye exam at the recommended PMA. However, that is the improvement opportunity.

Questions for the Committee:

- *Are the exclusions consistent with the evidence?*
- *Are any patients or patient groups inappropriately excluded from the measure? Are additional exclusions needed?*
- *Are the exclusions/exceptions of sufficient frequency and variation across providers to be needed (and outweigh the data collection burden)?*

2b4. Risk adjustment: **Risk-adjustment method** ☒ **None** ☐ **Statistical model** ☐ **Stratification**

2b5. Meaningful difference *(can statistically significant and clinically/practically meaningful differences in performance measure scores can be identified)*:

- The developers provide a table of results for each year from 2006-2014.
- The developer states that hospitals can compare their rates to others: "Members of Vermont Oxford Network can compare an individual hospital's unadjusted rate of retinopathy of prematurity screening to unadjusted mean for the Network as well as the unadjusted rates of the hospitals at the 25th and 75th percentile for the Network and by subgroup."

Question for the Committee:

- *Does this measure identify meaningful differences about quality?*

2b6. Comparability of data sources/methods:

N/A

2b7. Missing Data

- The developer states "Over the nine years, 947 infants (0.2%) were missing a value for retinal examination. Missing data does not appear to be an issue."

Guidance from algorithm #3 Evaluating Validity: Specifications consistent with evidence (Box 1) → Assessed all potential threats to validity (Box 2) → empirical testing (Box 3) →testing of measure score (Box 6) →appropriate testing method (Box 7) → testing results moderate → Rate as moderate

Question for the Committee:

- *Since the developer does not provide any new validity testing aside from updating the risk model discrimination, does the Committee accept the prior evaluation on this criterion without need for further discussion and voting?*

Preliminary rating for validity: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 2: Scientific Acceptability of Measure Properties (including all 2a, 2b, and 2d)

2a1. & 2b1. Specifications

Comments:

I am assuming that all babies born within this gestational age are registered, and therefore the denominator is accurate. How does the developer get the information on who has been screen (numerator)? Is this also from the registry or from charts???

If the facility is not a VON center, would you need to exclude transfer patients?

*Moderate**

2a2. Reliability Testing

Comments:

2b2. Validity Testing
Comments:

2b3. Exclusions Analysis
2b4. Risk Adjustment/Stratification for Outcome or Resource Use Measures
2b5. Identification of Statistically Significant & Meaningful Differences In Performance
2b6. Comparability of Performance Scores When More Than One Set of Specifications
2b7. Missing Data Analysis and Minimizing Bias
Comments:

Criterion 3. Feasibility
Maintenance measures – no change in emphasis – implementation issues may be more prominent

3. Feasibility is the extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

- All data is in electronic format in a clinical registry.
- There are no fees to use the measure, however, members of the Vermont Oxford Network pay an annual membership fee.

Questions for the Committee:

- *Is this registry-based measure available for use by everyone?*
- *Are the required data elements routinely generated and used during care delivery?*
- *Are the required data elements available in electronic form, e.g., EHR or other electronic sources?*
- *Is the data collection strategy ready to be put into operational use?*

Preliminary rating for feasibility: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 3: Feasibility

3a. Byproduct of Care Processes
3b. Electronic Sources
3c. Data Collection Strategy
Comments:
The is very feasible, if all the information comes from a registry to which all hospital must report. The question about how to insure that we are getting the correct data on which babies are test/not tested is still not clear to me.
May require some manual extraction to obtain the retinal exam. Registry is not available to everyone.

Criterion 4: Usability and Use
Maintenance measures – increased emphasis – much greater focus on measure use and usefulness, including both impact /improvement and unintended consequences

4. Usability and Use evaluate the extent to which audiences (e.g., consumers, purchasers, providers, policymakers) use or could use performance results for both accountability and performance improvement activities.

Current uses of the measure
Publicly reported? ☐ **Yes** ☒ **No**

Current use in an accountability program? ☐ **Yes** ☒ **No**
OR

Planned use in an accountability program? ☐ **Yes** ☒ **No**

Accountability program details

- This measure has been NQF endorsed since 2007. NQF criteria for usability and use is looking for "performance results are used in at least 1 accountability application within 3 years after initial endorsement and are publicly reported within 6 years after initial endorsement (or the data on performance results are available).
- The measure is currently used only for internal QI within the membership of VON. The developer does not know how many of the nearly 1000 members are using the measure.
- The developer states "Vermont Oxford Network is committed to working with accrediting bodies that are developing publicly-reported quality measures for the neonatal population."

Improvement results

- The proportion of infants receiving a timely retinopathy of prematurity examination has increased slightly from 2006 (90.1%) to 2014 (91.8%). The rate for hospitals at the 25th percentile increased from 88.2% in 2006 to 91.7% in 2012, decreasing to 90% in 2014.

Unexpected findings (positive or negative) during implementation

The developers are unaware of any unexpected findings or consequences. They provide detailed information to members on how to use the measure.

Potential harms: none reported

Feedback: none

***Questions for the Committee**:*

- *Despite being endorsed since 2007 this measure is neither publicly reported nor used in an accountability program. Does the Committee think that either is likely in the near future?*
- *How can the performance results be used to further the goal of high-quality, efficient healthcare?*
- *Do the benefits of the measure outweigh any potential unintended consequences?*

Preliminary rating for usability and use: ☐ **High** ☐ **Moderate** ☒ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 4: Usability and Use

4a. Accountability and Transparency
4b. Improvement
4c. Unintended Consequences

Comments:

**The screening is very important as it dictates the future well being of the premature infants for whom early intervention is paramount for this disease. **
Quality measure by VON only

Criterion 5: Related and Competing Measures

Related or competing measures

There are no related/competing measures.

Harmonization

N/A

Pre-meeting public and member comments

-

1. IMPACT, OPPORTUITY, EVIDENCE - IMPORTANCE TO MEASURE AND REPORT

Importance to Measure and Report is a threshold criterion that must be met in order to recommend a measure for endorsement. All three subcriteria must be met to pass this criterion. See guidance on evidence.
Measures must be judged to be important to measure and report in order to be evaluated against the remaining criteria. (evaluation criteria)

1c.1 Structure-Process-Outcome Relationship *(Briefly state the measure focus, e.g., health outcome, intermediate clinical outcome, process, structure; then identify the appropriate links, e.g., structure-process-health outcome; process- health outcome; intermediate clinical outcome-health outcome)*:
Health outcome: visual acuity, blindness
Intermediate clinical outcome: timely retinal ablation surgery
Process: processes for identifying eligible infants and scheduling and performing retinal exams
Structure: staffing with pediatric retinal specialists or other ophthalmology personnel

1c.2-3 Type of Evidence *(Check all that apply)*:
Clinical Practice Guideline
Selected individual studies (rather than entire body of evidence)
Systematic review of body of evidence (other than within guideline development)

1c.4 Directness of Evidence to the Specified Measure *(State the central topic, population, and outcomes addressed in the body of evidence and identify any differences from the measure focus and measure target population)*:
As far as we are aware the guidelines for timing of retinal examination have not been tested in a randomized trial. However, clear evidence from randomized trials on the effectiveness of surgical or pharmacological intervention for infants with threshold retinopathy of prematurity (ROP) leaves little doubt that appropriately timed retinal exams are required to identify infants who will benefit from treatment.

1c.5 Quantity of Studies in the Body of Evidence *(Total number of studies, not articles)*: Multiple studies cited below

1c.6 Quality of Body of Evidence (*Summarize the certainty or confidence in the estimates of benefits and harms to patients across studies in the body of evidence resulting from study factors. Please address: a) study design/flaws; b) directness/indirectness of the evidence to this measure (e.g., interventions, comparisons, outcomes assessed, population included in the evidence); and c) imprecision/wide confidence intervals due to few patients or events):* Strong evidence for the benefit of surgical or pharmacological intervention for infants with threshold ROP.

1c.7 Consistency of Results across Studies *(Summarize the consistency of the magnitude and direction of the effect):* Consistent

1c.8 Net Benefit *(Provide estimates of effect for benefit/outcome; identify harms addressed and estimates of effect; and net benefit - benefit over harms)*:
See summary of evidence above for improved visual outcomes

1c.9 Grading of Strength/Quality of the Body of Evidence. Has the body of evidence been graded? No

1c.10 If body of evidence graded, identify the entity that graded the evidence including balance of representation and any disclosures regarding bias: N/A

1c.11 System Used for Grading the Body of Evidence: Other

1c.12 If other, identify and describe the grading scale with definitions: Not graded

1c.13 Grade Assigned to the Body of Evidence: N/A

1c.14 Summary of Controversy/Contradictory Evidence: None cited

1c.15 Citations for Evidence other than Guidelines*(Guidelines addressed below)*:

Andersen C, Phelps D. Peripheral retinal ablation for threshold retinopathy of prematurity in preterm infants. Cochrane Database of Systematic Reviews 1999, Issue 3. Art. No.: CD001693. DOI: 10.1002/14651858.CD001693

Reynolds JD, Dobson V, Quinn GE, Fielder AR, Palmer EA, Saunders RA, Hardy RJ, Phelps DL, Baker JD, Trese MT, Schaffer DB, Tung B, for the CRYO-ROP and LIGHT-ROP Cooperative Groups: Evidence-based screening criteria for retinopathy of prematurity: natural history data from the CRYO-ROP and LIGHT-ROP studies. Arch Ophthalmol 120: 1470-1476, 2002.

Harrell SN, Brandon DH. Retinopathy of prematurity: the disease process, classifications, screening, treatment, and outcomes. Neonatal Netw 26(6): 371-378, 2007.

Salvin JH, Lehman SS, Jin J, Hendricks DH. Update on retinopathy of prematurity: treatment options and outcomes. Curr Opin Ophthalmol 21(5): 329-334, 2010.

Kennedy KA, Wrage LA, Higgins RD, Finer NN, Carlo WA, Walsh MC, Laptook AR, Faix RG, Yoder BA, Schibler K, Gantz MG, Das A, Newman NS, Phelps DL; SUPPORT Study Group of the Eunice Kennedy Shriver National Institute of Child Health and Human Development Neonatal Research Network. Evaluating retinopathy of prematurity screening guidelines for 24- to 27-week gestational age infants. J Perinatol 34(4): 311-318, 2014.

Taranath DA, Oh DD, Keane MC, Fabel H, Marshall P. Adequacy of published screening criteria for retinopathy of prematurity. Clin Experiment Ophthalmol 2015 Aug 7. 2015 Aug 7. doi: 10.1111/ceo.12628. [Epub ahead of print]

Kemper AR, Prosser LA, Wade KC, Repka MX, Ying GS, Baumritter A, Quinn GE; e-ROP Study Cooperative Group. A comparison of strategies for retinopathy of prematurity detection. Pediatrics 137(1): 1-10, 2016.

1c.16 Quote verbatim, <u>the specific guideline recommendation</u> *(Including guideline # and/or page #)*:

This statement revises a previous statement on screening of preterm infants for retinopathy of prematurity (ROP) that was published in 2006. ROP is a pathologic process that occurs only in immature retinal tissue and can progress to a tractional retinal detachment, which can result in functional or complete blindness. Use of peripheral retinal ablative therapy by using laser photocoagulation for nearly 2 decades has resulted in a high probability of markedly decreasing the incidence of this poor visual outcome, but the sequential nature of ROP creates a requirement that at-risk preterm infants be examined at proper times and intervals to detect the changes of ROP before they become permanently destructive. This statement presents the attributes on which an effective program for detecting and treating ROP could be based, including the timing of initial examination and subsequent reexamination intervals.

Infants with a birth weight of ≤1500 g or gestational age of 30 weeks or less (as defined by the attending neonatologist) and selected infants with a birth weight between 1500 and 2000 g or gestational age of >30 weeks with an unstable clinical course, including those requiring cardiorespiratory support and who are believed by their attending pediatrician or neonatologist to be at high risk for ROP, should have retinal screening examinations performed after pupillary dilation by using binocular indirect ophthalmoscopy with a lid speculum and scleral depression (as needed) to detect ROP. Dilating drops should be sufficient to allow adequate examination of the fundi, but care should be used in using multiple drops if the pupil fails to dilate, because poor pupillary dilation can occur in advanced ROP, and administering multiple doses of dilating drops can adversely affect the systemic status of the infant. Sterile instruments should be used to examine each infant to avoid possible cross-contamination of infectious agents. One examination is sufficient only if it unequivocally shows the retina to be fully vascularized in both eyes. Effort may be made to minimize the discomfort and systemic effect of this examination by pretreatment of the eyes with a topical anesthetic agent such as proparacaine; consideration also may be

given to the use of pacifiers, oral sucrose, and so forth.

Retinal examinations in preterm infants should be performed by an ophthalmologist who has sufficient knowledge and experience to identify accurately the location and sequential retinal changes of ROP. The International Classification of Retinopathy of Prematurity Revisited should be used to classify, diagram, and record these retinal findings at the time of examination.

The initiation of acute-phase ROP screening should be based on the infant's postmenstrual age. The onset of serious ROP correlates better with postmenstrual age (gestational age at birth plus chronologic age) than with postnatal age. That is, the more preterm an infant is at birth, the longer the time to develop serious ROP. This knowledge has been used previously in developing a screening schedule. Table 1 was developed from an evidence-based analysis of the Multicenter Trial of Cryotherapy for Retinopathy of Prematurity natural history data and confirmed by the Light Reduction in ROP Study, which was conducted a decade later. It represents a suggested schedule for the timing of the initial eye examinations based on postmenstrual age and chronologic (postnatal) age to detect ROP before it becomes severe enough to result in retinal detachment while minimizing the number of potentially traumatic examinations. Although Table 1 provides a schedule for detecting ROP potentially damaging to the retina with 99% confidence, it should be appreciated that infants born before 25 weeks' gestational age should be considered for earlier screening on the basis of severity of comorbidities (6 weeks' chronologic age, even if before 31 weeks' postmenstrual age to enable earlier identification and treatment of aggressive posterior ROP [a severe form of ROP that is characterized by rapid progression to advanced stages in posterior ROP] that is more likely to occur in this extremely high-risk population).

1-Week or Less Follow-up

immature vascularization: zone I—no ROP

immature retina extends into posterior zone II, near the boundary of zone I

stage 1 or 2 ROP: zone I
stage 3 ROP: zone II

the presence or suspected presence of aggressive posterior ROP

1- to 2-Week Follow-up

immature vascularization; posterior zone II

stage 2 ROP: zone II

unequivocally regressing ROP: zone I

2-Week Follow-up

stage 1 ROP: zone II

immature vascularization: zone II—no ROP

unequivocally regressing ROP: zone II

2- to 3-Week Follow-up

stage 1 or 2 ROP: zone III

regressing ROP: zone III

The conclusion of acute retinal screening examinations should be based on age and retinal ophthalmoscopic findings. Findings that suggest that examinations can be terminated include the following:

zone III retinal vascularization attained without previous zone I or II ROP (if there is examiner doubt about the zone or if the postmenstrual age is less than 35 weeks, confirmatory examinations may be warranted);

full retinal vascularization in close proximity to the ora serrata for 360°—that is, the normal distance found in mature retina between the end of vascularization and the ora serrata. This criterion should be used for all cases treated for ROP solely with bevacizumab;

postmenstrual age of 50 weeks and no prethreshold disease (defined as stage 3 ROP in zone II, any ROP in zone I) or worse ROP is present; or

regression of ROP (care must be taken to be sure that there is no abnormal vascular tissue present that is capable of reactivation and progression in zone II or III).

The use of digital photographic retinal images that are captured and sent for remote interpretation is a developing approach to ROP screening; however, outcomes comparison between large-scale operational digital-imaging systems with remote interpretation versus binocular indirect ophthalmoscopy have not been published. Nevertheless, some neonatal centers are conducting remote ROP screening for infants still in the hospital. At a minimum, programs that employ this method should comply with the timing and other recommendations outlined in the preceding guidelines. Protocol modifications may be required to allow for additional time for communication, processing, transportation, or other logistical issues. Captured images and their interpretation should be incorporated into the permanent medical record. It is also recommended that indirect ophthalmoscopy be performed at least once by a qualified ophthalmologist before treatment or termination of acute phase screening of ROP for infants at risk for ROP.

Digital image capture (taking of photographs) requires skill, experience, practice, a broad understanding of the infant eye, and ideally, a knowledge of the pathophysiology of ROP (zone, stage, and plus). Remote ROP graders should have the same training requirements as bedside examiners and a mentored experience in interpretation of digital images for ROP. Interpretation requires not only expert knowledge about ROP but also understanding of the limitations of interpreting static images and the special care that must be taken to schedule more frequent imaging sessions that may be required because of those limitations. Remote interpreters must provide clinical input on the timing of follow-up imaging sessions and ophthalmoscopic examinations and appropriate methodology, and these findings need to be communicated in a manner that is compliant with rules of the Health Insurance Portability and Accountability Act (HIPAA).

Digital retinal imaging may also be a useful tool for objective documentation of retinal findings and for teaching NICU staff and parents about examination results, even if it is not the primary method used for ROP screening in the NICU.

ROP care that includes off-site photographic interpretation requires close collaboration among neonatologists, imaging staff, and ophthalmologists. As with all ROP screening programs, specific responsibilities of each individual must be carefully delineated in a written protocol in advance so that repeat imaging and/or confirmatory examinations and required treatments can be performed without delay.

1c.17 **Clinical Practice Guideline Citation:** American Academy of Pediatrics, Section on Ophthalmology, American Academy of Ophthalmology, American Association for Pediatric Ophthalmology and Strabismus. Screening examination of premature infants for retinopathy of prematurity. Pediatrics. 2013; 131 :189-195.

1c.18 **National Guideline Clearinghouse or other URL:**

1c.19 **Grading of Strength of Guideline Recommendation.** Has the recommendation been graded? No

1c.20 **If guideline recommendation graded, identify the entity that graded the evidence including balance of representation and any disclosures regarding bias:**

1c.21 **System Used for Grading the Strength of Guideline Recommendation:** Other

1c.22 **If other, identify and describe the grading scale with definitions:** Not graded

1c.23 **Grade Assigned to the Recommendation:** Not graded

1c.24 **Rationale for Using this Guideline Over Others:** AAP is the authoritative source.
Based on the NQF descriptions for rating the evidence, what was the <u>developer's assessment</u> of the quantity, quality, and consistency of the body of evidence? 1c.25 Quantity: High 1c.26 Quality: High1c.27 Consistency: High

1. Evidence, Performance Gap, Priority – Importance to Measure and Report

Extent to which the specific measure focus is evidence-based, important to making significant gains in healthcare quality, and improving health outcomes for a specific high-priority (high-impact) aspect of healthcare where there is variation in or overall less-than-optimal performance. ***Measures must be judged to meet all subcriteria to pass this criterion and be evaluated against the remaining criteria.***

1a. Evidence to Support the Measure Focus – See attached Evidence Submission Form
ROP_evidence-635930518744149557.doc

1b. Performance Gap

Demonstration of quality problems and opportunity for improvement, i.e., data demonstrating:

- considerable variation, or overall less-than-optimal performance, in the quality of care across providers; and/or
- disparities in care across population groups.

1b.1. **Briefly explain the rationale for this measure** *(e.g., the benefits or improvements in quality envisioned by use of this measure)*
The goal of this measure is to increase proportion of high risk infants who receive retinal screening on the recommended schedule.

1b.2. **Provide performance scores on the measure as specified (current and over time) at the specified level of analysis.** *(This is required for endorsement maintenance. Include mean, std dev, min, max, interquartile range, scores by decile. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included). This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*
In 2014, 916 hospitals in the Vermont Oxford Network enrolled 40,638 eligible infants with gestational age between 22 and 29 weeks, and 32,325 (79.5%) were in the reporting hospital at the postnatal age recommended for ROP screening. Of those, 6.2% did not receive retinal exams prior to discharge, with the interquartile range among hospitals of 0% to 10% not receiving recommended screenings.

Proportion of Eligible Infants Receiving ROP Screening, by Center and Year:

Year	N	Mean	Min	Max	10th Pctl	25th Pctl	75th Pctl	90th Pctl
2006	633	0.901	0.000	1.000	0.781	0.882	1.000	1.000
2007	690	0.908	0.000	1.000	0.800	0.889	1.000	1.000
2008	762	0.917	0.000	1.000	0.800	0.909	1.000	1.000
2009	803	0.914	0.000	1.000	0.786	0.909	1.000	1.000
2010	848	0.919	0.000	1.000	0.800	0.907	1.000	1.000
2011	872	0.929	0.000	1.000	0.824	0.917	1.000	1.000
2012	886	0.924	0.000	1.000	0.800	0.917	1.000	1.000
2013	900	0.919	0.000	1.000	0.800	0.909	1.000	1.000
2014	916	0.918	0.000	1.000	0.786	0.900	1.000	1.000

1b.3. **If no or limited performance data on the measure as specified is reported in 1b2, then provide a summary of data from the literature that indicates opportunity for improvement or overall less than optimal performance on the specific focus of measurement.**
From 2000 to 2009, rates of any retinal examination among infants weighing 501 to 1500 g and cared for at Vermont Oxford Network members in North America increased from 66.2% to 74.5% (P value < 0.001). (Soll RF, Edwards EM, Badger GJ, et al. Obstetric and neonatal care practices for infants 501 to 1500 g from 2000 to 2009. Pediatrics 2013;132(2): 222-228). In a survey of neonatologists, 29% agreed that some children in their state develop ROP-related visual impairment that could have been prevented with timely screening (Kemper AR, Wallace DK. Neonatologists´ practices and experiences in arranging retinopathy of prematurity screening services. Pediatrics 2007;120:527-531).

2. American Academy of Pediatrics, Section on Ophthalmology, American Academy of Ophthalmology, American Association for Pediatric Ophthalmology and Strabismus. Errata: screening examination of premature infants for retinopathy of prematurity. Pediatrics. 2006;118 :132.

3. Kemper AR, Wallace DK. Neonatologists´ Practices and Experiences in Arranging Retinopathy of Prematurity Screening Services. Pediatrics 2007;120:527-531.

4. Vermont Oxford Network unbublished data, 2007.

1b.4. **Provide disparities data from the measure as specified (current and over time) by population group, e.g., by race/ethnicity, gender, age, insurance status, socioeconomic status, and/or disability.** *(This is required for endorsement maintenance. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities include.) This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*
In 2014, there were no differences by maternal race in rates of eye examination at the recommended age: White non-Hispanic (95%); black non-Hispanic (93%); hispanic (95%); Asian (94%).

Proportion of Eligible Infants Receiving ROP Screening:

	White non-Hispanic		Black non-Hispanic	Hispanic Asian
2006	0.927	0.926	0.931	0.938
2007	0.932	0.921	0.931	0.914
2008	0.935	0.923	0.935	0.939
2009	0.942	0.928	0.940	0.936
2010	0.944	0.934	0.940	0.937
2011	0.947	0.943	0.948	0.945
2012	0.946	0.932	0.952	0.949
2013	0.942	0.930	0.954	0.936
2014	0.946	0.925	0.948	0.939

1b.5. **If no or limited data on disparities from the measure as specified is reported in 1b4, then provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement. Include citations.**
N/A

1c. High Priority (previously referred to as High Impact)

The measure addresses:

- a specific national health goal/priority identified by DHHS or the National Priorities Partnership convened by NQF; OR
- a demonstrated high-priority (high-impact) aspect of healthcare (e.g., affects large numbers of patients and/or has a substantial impact for a smaller population; leading cause of morbidity/mortality; high resource use (current and/or future); severity of illness; and severity of patient/societal consequences of poor quality).

1c.1. **Demonstrated high priority aspect of healthcare**
Affects large numbers, A leading cause of morbidity/mortality, Frequently performed procedure, Patient/societal consequences of poor quality
1c.2. **If Other:**

1c.3. **Provide epidemiologic or resource use data that demonstrates the measure addresses a high priority aspect of healthcare. List citations in 1c.4.**
The clinical guideline states: "Retinopathy of prematurity (ROP) is a disorder of the developing retina of low birth weight preterm infants that potentially leads to blindness in a small but significant percentage of those infants. ...Because of the sequential nature of ROP progression and the proven benefits of timely treatment in reducing the risk of visual loss, effective care now requires that at-risk infants receive carefully timed retinal examinations by an ophthalmologist who is experienced in the examination of preterm infants for ROP using a binocular indirect opthalmoscope on a scheduled basis according to their gestational age at birth." In 2000, 10.5% of infants weighing 501 to 1500 g and cared for at Vermont Oxford Network members in North America developed severe ROP, decreasing to 6.8% in 2009 (P value < 0.001).

1c.4. **Citations for data demonstrating high priority provided in 1a.3**
American Academy of Pediatrics, Section on Ophthalmology, American Academy of Ophthalmology, American Association for Pediatric Ophthalmology and Strabismus. Screening examination of premature infants for retinopathy of prematurity. Pediatrics. 2013; 131: 189-195.

Horbar, JD, Carpenter JH, Badger GJ, et al. Mortality and neonatal morbidity among infants 501 to 1500 grams from 2000 to 2009. Pediatrics. 2012; 129(6): 1019-1026.

1c.5. **If a PRO-PM (e.g. HRQoL/functional status, symptom/burden, experience with care, health-related behaviors), provide evidence that the target population values the measured PRO and finds it meaningful.** *(Describe how and from whom their input*

was obtained.)

2. Reliability and Validity—Scientific Acceptability of Measure Properties

Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. ***Measures must be judged to meet the subcriteria for both reliability and validity to pass this criterion and be evaluated against the remaining criteria.***

2a.1. Specifications The measure is well defined and precisely specified so it can be implemented consistently within and across organizations and allows for comparability. eMeasures should be specified in the Health Quality Measures Format (HQMF) and the Quality Data Model (QDM).

De.5. **Subject/Topic Area** *(check all the areas that apply):*
Perinatal and Reproductive Health

De.6. **Cross Cutting Areas** *(check all the areas that apply):*
Safety : Complications

S.1. **Measure-specific Web Page** *(Provide a URL link to a web page specific for this measure that contains current detailed specifications including code lists, risk model details, and supplemental materials. Do not enter a URL linking to a home page or to general information.)*
https://public.vtoxford.org/databases/nqf-endorsed-measures/

S.2a. **If this is an eMeasure**, HQMF specifications must be attached. Attach the zipped output from the eMeasure authoring tool (MAT) - if the MAT was not used, contact staff. (Use the specification fields in this online form for the plain-language description of the specifications)
This is not an eMeasure **Attachment:**

S.2b. **Data Dictionary, Code Table, or Value Sets** *(and risk model codes and coefficients when applicable) must be attached. (Excel or csv file in the suggested format preferred - if not, contact staff)*
Attachment **Attachment:** 0483_ICD.xlsx

S.3. **For endorsement maintenance**, please briefly describe any changes to the measure specifications since last endorsement date and explain the reasons.
No changes

S.4. **Numerator Statement** *(Brief, narrative description of the measure focus or what is being measured about the target population, i.e., cases from the target population with the target process, condition, event, or outcome)*
IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
Number of infants born from 22 weeks, 0 days to 29 weeks, 6 days gestational age who were in the reporting hospital at the postnatal age recommended for ROP screening by the AAP and who received a retinal exam for ROP prior to discharge

S.5. **Time Period for Data** *(What is the time period in which data will be aggregated for the measure, e.g., 12 mo, 3 years, look back to August for flu vaccination? Note if there are different time periods for the numerator and denominator.)*
Twelve month period (Jan 1 - Dec 31)

S.6. **Numerator Details** *(All information required to identify and calculate the cases from the target population with the target process, condition, event, or outcome such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
IF an OUTCOME MEASURE, describe how the observed outcome is identified/counted. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
Number of infants born from 22 weeks, 0 days to 29 weeks, 6 days gestational age who were in the reporting hospital at the postnatal age recommended for ROP screening by the AAP and who received a retinal exam for ROP prior to discharge

S.7. **Denominator Statement** *(Brief, narrative description of the target population being measured)*
All eligible infants born from 22 weeks, 0 days to 29 weeks, 6 days gestational age who were in the reporting hospital at the postnatal age recommended for ROP screening by the AAP

S.8. **Target Population Category** *(Check all the populations for which the measure is specified and tested if any):*

Children's Health

S.9. **Denominator Details** *(All information required to identify and calculate the target population/denominator such as definitions, specific data collection items/responses , code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
Any infant who is born at the reporting hospital and whose gestational age at birth is
from 22 weeks, 0 days to 29 weeks, 6 days should be included if they are in the reporting hospital at the postnatal age recommended for ROP screening by the AAP.
Any outborn infant who is admitted to any location in the reporting hospital within 28 days of birth, without first having gone home, and whose gestational age is from 22 weeks, 0 days to 29 weeks, 6 days should be included if they are in the reporting hospital at the postnatal age recommended for ROP screening by the AAP.

S.10. **Denominator Exclusions** *(Brief narrative description of exclusions from the target population)*
1. Infants outside the gestational age range of 22 to 29 weeks
2. Outborn infants admitted to the reporting hospital more than 28 days after birth
3. Outborn infants who have been home prior to admission
4. Infants who die in the delivery room or initial resuscitation area prior to admission to the neonatal intensive care unit
5. Infants not in the reporting hospital at the postnatal age recommended for ROP screening by the AAP

S.11. **Denominator Exclusion Details** *(All information required to identify and calculate exclusions from the denominator such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
See S.10. above.

S.12. **Stratification Details/Variables** *(All information required to stratify the measure results including the stratification variables, definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format with at S.2b)*
Reports are stratified by gestational age, birth location and birth weight category.

S.13. **Risk Adjustment Type** (Select type. Provide specifications for risk stratification in S.12 and for statistical model in S.14-15)
Stratification by risk category/subgroup
If other:

S.14. **Identify the statistical risk model method and variables** *(Name the statistical method - e.g., logistic regression and list all the risk factor variables. Note - risk model development and testing should be addressed with measure testing under Scientific Acceptability)*
N/A

S.15. **Detailed risk model specifications** *(must be in attached data dictionary/code list Excel or csv file. Also indicate if available at measure-specific URL identified in S.1.)*
Note: Risk model details (including coefficients, equations, codes with descriptors, definitions), should be provided on a separate worksheet in the suggested format in the Excel or csv file with data dictionary/code lists at S.2b.

S.15a. **Detailed risk model specifications** (*if not provided in excel or csv file at S.2b*)
N/A

S.16. **Type of score:**
Rate/proportion
If other:

S.17. **Interpretation of Score** *(Classifies interpretation of score according to whether better quality is associated with a higher score, a lower score, a score falling within a defined interval, or a passing score)*
Better quality = Higher score

S.18. **Calculation Algorithm/Measure Logic** *(Describe the calculation of the measure score as an ordered sequence of steps including identifying the target population; exclusions; cases meeting the target process, condition, event, or outcome; aggregating data; risk adjustment; etc.)*

1. Identify the population of eligible infants: all infants whose gestational age at birth is from 22 weeks, 0 days, to 29 weeks, 6 days, who are born at or admitted to the hospital within 28 days of birth without having been discharged home and who are still hospitalized at the postnatal age at which the first retinal screening exam is recommended by the AAP guidelines.

a. Determine the infant's postnatal age at discharge. This is calculated in days as date of discharge minus date of admission +1. Divide by 7 to determine the postnatal age at discharge in weeks.

b. Compare each infant's postnatal age at discharge to the appropriate row in the following table adapted from: American Academy of Pediatrics, Section on Ophthalmology, American Academy of Ophthalmology, American Association for Pediatric Ophthalmology and Strabismus. Screening examination of premature infants for retinopathy of prematurity. Pediatrics. 2013;131:189.

Timing of First Eye Examination Based on Gestational Age at Birth

Gestational age at birth (completed weeks)	Postnatal Age (weeks) at initial ROP screening exam
22	9
23	8
24	7
25	6
26	5
27	4
28	4
29	4

c. If the infant's postnatal age at discharge is greater than or equal to the postnatal age for initial ROP screening from the table, the infant is classified as "still hospitalized at the time of recommended initial ROP screening".

2. Among the population of eligible infants:

a. Count the number of infants in the population of eligible infants. This number is the denominator for the measure: DENOM.

b. Count the number of infants who had a retinal examination prior to discharge. This number is the numerator for the measure: NUM.

c. The measure is calculated as:
NUM / DENOM
This measure represents the proportion of infants 22 to 29 weeks gestation who were hospitalized at the age when ROP screening is recommended who were screened prior to discharge.

d. To stratify by gestational age, limit the counts and calculation to infants in the gestational age for the range 22-29 weeks.

S.19. **Calculation Algorithm/Measure Logic Diagram URL or Attachment** *(You also may provide a diagram of the Calculation Algorithm/Measure Logic described above at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
No diagram provided

S.20. **Sampling** *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*
IF a PRO-PM, identify whether (and how) proxy responses are allowed.
Data for all eligible infants born during the reporting period are collected.

S.21. **Survey/Patient-reported data** *(If measure is based on a survey, provide instructions for conducting the survey and guidance on minimum response rate.)*
IF a PRO-PM, specify calculation of response rates to be reported with performance measure results.

N/A

S.22. **Missing data** (specify how missing data are handled, e.g., imputation, delete case.)
Required for Composites and PRO-PMs.
Omit infants for whom information on screening is not known and unknowable.

S.23. **Data Source** *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*
If other, please describe in S.24.
Electronic Clinical Data : Registry

S.24. **Data Source or Collection Instrument** *(Identify the specific data source/data collection instrument e.g. name of database, clinical registry, collection instrument, etc.)*
IF a PRO-PM, identify the specific PROM(s); and standard methods, modes, and languages of administration.
Vermont Oxford Network Database

S.25. **Data Source or Collection Instrument** *(available at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
No data collection instrument provided

S.26. **Level of Analysis** *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*
Facility

S.27. **Care Setting** *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*
Hospital/Acute Care Facility
If other:

S.28. **COMPOSITE Performance Measure** - Additional Specifications *(Use this section as needed for aggregation and weighting rules, or calculation of individual performance measures if not individually endorsed.)*

2a. Reliability – See attached Measure Testing Submission Form
2b. Validity – See attached Measure Testing Submission Form
ROP_testing_attachment-635930519958952705.docx

Measure Number (*if previously endorsed*)**:** 483

Measure Title: Proportion of infants 22 to 29 weeks gestation screened for retinopathy of prematurity

Date of Submission: 2/5/2016

Type of Measure:

☐ Composite – ***STOP – use composite testing form***	☐ Outcome (*including PRO-PM*)
☐ Cost/resource	☒ Process
☐ Efficiency	☐ Structure

Instructions

- Measures must be tested for all the data sources and levels of analyses that are specified. ***If there is more than one set of data specifications or more than one level of analysis, contact NQF staff*** about how to present all the testing information in one form.
- **For all measures, sections 1, 2a2, 2b2, 2b3, and 2b5 must be completed.**
- **For outcome and resource use measures**, section **2b4** also must be completed.
- If specified for **multiple data sources/sets of specificaitons** (e.g., claims and EHRs), section **2b6** also must be completed.
- Respond to all questions as instructed with answers immediately following the question. All information on testing to demonstrate meeting the subcriteria for reliability (2a2) and validity (2b2-2b6) must be in this form. An appendix for *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 20 pages (*incuding questions/instructions;* minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.
- For information on the most updated guidance on how to address sociodemographic variables and testing in this form refer to the release notes for version 6.6 of the Measure Testing Attachment.

Note: The information provided in this form is intended to aid the Steering Committee and other stakeholders in understanding to what degree the testing results for this measure meet NQF's evaluation criteria for testing.

2a2. Reliability testing [10] demonstrates the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise. For **PRO-PMs and composite performance measures**, reliability should be demonstrated for the computed performance score.

2b2. Validity testing [11] demonstrates that the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality. For **PRO-PMs and composite performance measures**, validity should be demonstrated for the computed performance score.

2b3. Exclusions are supported by the clinical evidence; otherwise, they are supported by evidence of sufficient frequency of occurrence so that results are distorted without the exclusion; [12]
AND
If patient preference (e.g., informed decisionmaking) is a basis for exclusion, there must be evidence that the exclusion impacts performance on the measure; in such cases, the measure must be specified so that the information about patient preference and the effect on the measure is transparent (e.g., numerator category computed separately, denominator exclusion category computed separately). [13]

2b4. For outcome measures and other measures when indicated (e.g., resource use):

- **an evidence-based risk-adjustment strategy** (e.g., risk models, risk stratification) is specified; is based on patient factors (including clinical and sociodemographic factors) that influence the measured outcome and are present at start of care; [14,15] and has demonstrated adequate discrimination and calibration

OR

- rationale/data support no risk adjustment/ stratification.

2b5. Data analysis of computed measure scores demonstrates that methods for scoring and analysis of the specified measure allow for **identification of statistically significant and practically/clinically meaningful [16] differences in performance**;

OR

there is evidence of overall less-than-optimal performance.

2b6. If multiple data sources/methods are specified, there is demonstration they produce comparable results.

2b7. For **eMeasures, composites, and PRO-PMs** (or other measures susceptible to missing data), analyses identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias.

Notes

10. Reliability testing applies to both the data elements and computed measure score. Examples of reliability testing for data elements include, but are not limited to: inter-rater/abstractor or intra-rater/abstractor studies; internal consistency for multi-item scales; test-retest for survey items. Reliability testing of the measure score addresses precision of measurement (e.g., signal-to-noise).

11. Validity testing applies to both the data elements and computed measure score. Validity testing of data elements typically analyzes agreement with another authoritative source of the same information. Examples of validity testing of the measure score include, but are not limited to: testing hypotheses that the measures scores indicate quality of care, e.g., measure scores are different for groups known to have differences in quality assessed by another valid quality measure or method; correlation of measure scores with another valid indicator of quality for the specific topic; or relationship to conceptually related measures (e.g., scores on process measures to scores on outcome measures). Face validity of the measure score as a quality indicator may be adequate if accomplished through a systematic and transparent process, by identified experts, and explicitly addresses whether performance scores resulting from the measure as specified can be used to distinguish good from poor quality.

12. Examples of evidence that an exclusion distorts measure results include, but are not limited to: frequency of occurrence, variability of exclusions across providers, and sensitivity analyses with and without the exclusion.

13. Patient preference is not a clinical exception to eligibility and can be influenced by provider interventions.

14. Risk factors that influence outcomes should not be specified as exclusions

15. With large enough sample sizes, small differences that are statistically significant may or may not be practically or clinically meaningful. The substantive question may be, for example, whether a statistically significant difference of one percentage point in the percentage of patients who received smoking cessation counseling (e.g., 74 percent v. 75 percent) is clinically meaningful; or whether a statistically significant difference of $25 in cost for an episode of care (e.g., $5,000 v. $5,025) is practically meaningful. Measures with overall less-than-optimal performance may not demonstrate much variability across providers.

1. DATA/SAMPLE USED FOR ALL TESTING OF THIS MEASURE

Often the same data are used for all aspects of measure testing. In an effort to eliminate duplication, the first five questions apply to all measure testing. If there are differences by aspect of testing,(e.g., reliability vs. validity) be sure to indicate the specific differences in question 1.7.

1.1. What type of data was used for testing? (*Check all the sources of data identified in the measure specifications and data used for testing the measure*. *Testing must be provided for all the sources of data specified and intended for*

measure implementation. ***If different data sources are used for the numerator and denominator, indicate N [numerator] or D [denominator] after the checkbox.***)

Measure Specified to Use Data From: (*must be consistent with data sources entered in S.23*)	**Measure Tested with Data From:**
☐ abstracted from paper record	☐ abstracted from paper record
☐ administrative claims	☐ administrative claims
☒ clinical database/registry	☒ clinical database/registry
☐ abstracted from electronic health record	☐ abstracted from electronic health record
☐ eMeasure (HQMF) implemented in EHRs	☐ eMeasure (HQMF) implemented in EHRs
☐ other: Click here to describe	☐ other: Click here to describe

1.2. If an existing dataset was used, identify the specific dataset (*the dataset used for testing must be consistent with the measure specifications for target population and healthcare entities being measured; e.g., Medicare Part A claims, Medicaid claims, other commercial insurance, nursing home MDS, home health OASIS, clinical registry*). Vermont Oxford Network Database

1.3. What are the dates of the data used in testing? 2006 to 2014

1.4. What levels of analysis were tested? (*testing must be provided for <u>all</u> the levels specified and intended for measure implementation, e.g., individual clinician, hospital, health plan*)

Measure Specified to Measure Performance of: (*must be consistent with levels entered in item S.26*)	**Measure Tested at Level of:**
☐ individual clinician	☐ individual clinician
☐ group/practice	☐ group/practice
☒ hospital/facility/agency	☒ hospital/facility/agency
☐ health plan	☐ health plan
☐ other:	☐ other:

1.5. How many and which <u>measured entities</u> were included in the testing and analysis (by level of analysis and data source)? (*identify the number and descriptive characteristics of measured entities included in the analysis (e.g., size, location, type); if a sample was used, describe how entities were selected for inclusion in the sample*) 1,044 hospitals contributed at least one year of data.

1.6. How many and which <u>patients</u> were included in the testing and analysis (by level of analysis and data source)? (*identify the number and descriptive characteristics of patients included in the analysis (e.g., age, sex, race, diagnosis); if a sample was used, describe how patients were selected for inclusion in the sample*) Over the nine year period, 320,474 infants who were 22 to 29 weeks gestational age at birth who did not die in the delivery room or within 12 hours of birth were registered with Vermont Oxford Network of which 271,098 were eligible to have an eye exam based on postmenstrual age.

1.7. If there are differences in the data or sample used for different aspects of testing (e.g., reliability, validity, exclusions, risk adjustment), identify how the data or sample are different for each aspect of testing reported below. None

1.8 What were the patient-level sociodemographic (SDS) variables that were available and analyzed in the data or sample used? For example, patient-reported data (e.g., income, education, language), proxy variables when SDS data are not collected from each patient (e.g. census tract), or patient community characteristics (e.g. percent vacant housing, crime rate). None

---

2a2. RELIABILITY TESTING

***Note**: If accuracy/correctness (validity) of data elements was empirically tested, separate reliability testing of data elements is not required – in 2a2.1 check critical data elements; in 2a2.2 enter "see section 2b2 for validity testing of data elements"; and skip 2a2.3 and 2a2.4.*

2a2.1. What level of reliability testing was conducted? (*may be one or both levels*)
☐ **Critical data elements used in the measure** (*e.g., inter-abstractor reliability; data element reliability must address ALL critical data elements*)
☒ **Performance measure score** (e.g., *signal-to-noise analysis*)

2a2.2. For each level checked above, describe the method of reliability testing and what it tests (*describe the steps—do not just name a method; what type of error does it test; what statistical analysis was used*)
We did a split-half analysis to measure the internal consistency of the measure. We selected hospitals with at least 10 infants, ran 100 samples of patients within each hospital randomly split to two groups, computed hospital rates, and calculated the correlations.

2a2.3. For each level of testing checked above, what were the statistical results from reliability testing? (e.*g., percent agreement and kappa for the critical data elements; distribution of reliability statistics from a signal-to-noise analysis*)
Correlation coefficients are:

2006	0.74
2007	0.78
2008	0.79
2009	0.78
2010	0.75
2011	0.69
2012	0.69
2013	0.72
2014	0.72

2a2.4 What is your interpretation of the results in terms of demonstrating reliability? (i.*e., what do the results mean and what are the norms for the test conducted?*)
The correlation coefficients were lower than we expected. It suggests that the definition may not be applied in the same manner across all infants at all hospitals. The coefficients increased as the number of infants at the hospitals increased.

2b2. VALIDITY TESTING
2b2.1. What level of validity testing was conducted? (*may be one or both levels*)
☐ **Critical data elements** (*data element validity must address ALL critical data elements*)
☐ **Performance measure score**
  ☐ **Empirical validity testing**
  ☒ **Systematic assessment of face validity of performance measure score as an indicator** of quality or resource use (*i.e., is an accurate reflection of performance on quality or resource use and can distinguish good from poor performance*)

2b2.2. For each level of testing checked above, describe the method of validity testing and what it tests (*describe the steps—do not just name a method; what was tested, e.g., accuracy of data elements compared to authoritative source, relationship to another measure as expected; what statistical analysis was used)* The measure was developed by board

certified neonatologists and reviewed by clinical experts in pediatric ophthalmology. The measure is reviewed annually by the Vermont Oxford Network Database Advisory Committee, consisting of national and international experts in the neonatal community. Comprehensive business rules test reach record for consistency, completeness and accuracy. Submitted records with errors must be corrected before data are finalized and reports of the measure are provided to hospitals.

2b2.3. What were the statistical results from validity testing? (*e.g., correlation; t-test*) No testing done

2b2.4. What is your interpretation of the results in terms of demonstrating validity? (i.*e., what do the results mean and what are the norms for the test conducted?*) N/A

2b3. EXCLUSIONS ANALYSIS
NA ☐ no exclusions — *skip to section 2b4*

2b3.1. Describe the method of testing exclusions and what it tests (*describe the steps—do not just name a method; what was tested, e.g., whether exclusions affect overall performance scores; what statistical analysis was used*)

Infants can be excluded in two ways. First, we do not require hospitals to determine if a transferred infant received a retinal examination unless the infant was readmitted. We would expect the percentage of infants who received a retinal examination to be low if infants are transferred early. We examined the correlation coefficient at the hospital level between the proportion of infants who were transferred by day 3 after birth and the proportion who had an eye examination at the recommended post menstrual age. The null hypothesis was that is a negative correlation between these two variables.

Second, infants may have had retinal examinations but not at the correct postmenstrual age. We examined the correlation coefficient at the hospital level between the proportion of infants who had a retinal examination and the proportion of infants who had an examination at the recommended post menstrual age. The null hypothesis was that there is a strong, positive correlation between these two variables.

2b3.2. What were the statistical results from testing exclusions? (*include overall number and percentage of individuals excluded, frequency distribution of exclusions across measured entities, and impact on performance measure scores*)

The correlation coefficient between the proportion of early transfers and the proportion who received a retinal exam at the recommended post menstrual age was -0.124 (P<0.0001). However, the average proportion of early transfers was 0.72%; over half of the hospitals had an early transfer rate of 0%. Therefore it is difficult to judge the effect of early transfer on eye exams by the size of the correlation coefficient. Among hospitals that had any early transfers, the correlation coefficient was -0.186 (P<0.0001).

The correlation coefficient between the proportion of infants who had a retinal examination and the proportion who had one at the recommended post menstrual age was 0.833 (P<0.0001). The difference between the proportion of infants who had a retinal examination and those who had one at the recommended post menstrual age was 12% on average, ranging from 0% (all infants had eye exams at the recommended PMA) to 89% who did not have the eye exam at the recommended PMA. However, that is the improvement opportunity.

2b3.3. What is your interpretation of the results in terms of demonstrating that exclusions are needed to prevent unfair distortion of performance results? (*i.e., the value outweighs the burden of increased data collection and analysis. Note: **If patient preference is an exclusion**, the measure must be specified so that the effect on the performance score is transparent, e.g., scores with and without exclusion*) The Vermont Oxford Network reporting has both the percent of infants who received any retinal examination and the percent who received an examination at the recommended post menstrual age. We encourage centers to use both measures to identify improvement opportunities. If hospitals transfer

infants early to higher levels of care, a different quality measure may be more appropriate.

2b4. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES
If not an intermediate or health outcome, or PRO-PM, or resource use measure, skip to section 2b5.

2b4.1. What method of controlling for differences in case mix is used?
☒ **No risk adjustment or stratification**
☐ **Statistical risk model with** Click here to enter number of factors **risk factors**
☐ **Stratification by** Click here to enter number of categories **risk categories**
☐ **Other,** Click here to enter description

2b4.2. If an outcome or resource use measure is not risk adjusted or stratified, provide rationale and analyses to demonstrate that controlling for differences in patient characteristics (case mix) is not needed to achieve fair comparisons across measured entities. Whether an infant receives screening for retinopathy of prematurity is unlikely to be different across patient characteristics.

2b4.3. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors (clinical factors or sociodemographic factors) used in the statistical risk model or for stratification by risk (*e.g., potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of p<0.10; correlation of x or higher; patient factors should be present at the start of care*) N/A

2b4.4a. What were the statistical results of the analyses used to select risk factors?

2b4.4b. Describe the analyses and interpretation resulting in the decision to select SDS factors (e.g. prevalence of the factor across measured entities, empirical association with the outcome, contribution of unique variation in the outcome, assessment of between-unit effects and within-unit effects)

2b4.5. Describe the method of testing/analysis used to develop and validate the adequacy of the statistical model or stratification approach (*describe the steps—do not just name a method; what statistical analysis was used*)

Provide the statistical results from testing the approach to controlling for differences in patient characteristics (case mix) below.
If stratified, skip to 2b4.9

2b4.6. Statistical Risk Model Discrimination Statistics (*e.g., c-statistic, R-squared*)**:**

2b4.7. Statistical Risk Model Calibration Statistics (*e.g., Hosmer-Lemeshow statistic*):

2b4.8. Statistical Risk Model Calibration – Risk decile plots or calibration curves:

2b4.9. Results of Risk Stratification Analysis:

2b4.10. What is your interpretation of the results in terms of demonstrating adequacy of controlling for differences in patient characteristics (case mix)? (i.*e., what do the results mean and what are the norms for the test conducted*)

2b4.11. Optional Additional Testing for Risk Adjustment (*not required, but would provide additional support of adequacy of risk model, e.g., testing of risk model in another data set; sensitivity analysis for missing data; other methods that were assessed*)

2b5. IDENTIFICATION OF STATISTICALLY SIGNIFICANT & MEANINGFUL DIFFERENCES IN PERFORMANCE

2b5.1. Describe the method for determining if statistically significant and clinically/practically meaningful differences in performance measure scores among the measured entities can be identified (*describe the steps—do not just name a method; what statistical analysis was used? Do not just repeat the information provided related to performance gap in 1b)* Members of Vermont Oxford Network can compare an individual hospital's unadjusted rate of retinopathy of prematurity screening to unadjusted mean for the Network as well as the unadjusted rates of the hospitals at the 25th and 75th percentile for the Network and by subgroup.

2b5.2. What were the statistical results from testing the ability to identify statistically significant and/or clinically/practically meaningful differences in performance measure scores across measured entities? (e.g., *number and percentage of entities with scores that were statistically significantly different from mean or some benchmark, different from expected; how was meaningful difference defined*)
The number and percent of eligible infants who were in the reporting hospital at the postnatal age recommended for retinopathy of prematurity screening by the American Academy of Pediatrics and who received a retinal examination prior to discharge are reported, along with mean values and the 25th and 75th percentile values for all hospitals.

Year	N hospitals	N eligible infants	Mean	25th Pctl	75th Pctl
2006	633	22,202	0.901	0.882	1.000
2007	690	28,364	0.908	0.889	1.000
2008	762	30,290	0.917	0.909	1.000
2009	803	30,823	0.914	0.909	1.000
2010	848	30,833	0.919	0.907	1.000
2011	872	31,572	0.929	0.917	1.000
2012	886	32,103	0.924	0.917	1.000
2013	900	32,132	0.919	0.909	1.000
2014	916	32,782	0.918	0.900	1.000

2b5.3. What is your interpretation of the results in terms of demonstrating the ability to identify statistically significant and/or clinically/practically meaningful differences in performance across measured entities? (i.*e., what do the results mean in terms of statistical and meaningful differences?*) N/A

2b6. COMPARABILITY OF PERFORMANCE SCORES WHEN MORE THAN ONE SET OF SPECIFICATIONS
***If only one set of specifications, this section can be skipped*.**

Note: *This item is directed to measures that are risk-adjusted (with or without SDS factors)* ***OR*** *to measures with more than one set of specifications/instructions (e.g., one set of specifications for how to identify and compute the measure from medical record abstraction and a different set of specifications for claims or eMeasures). It does not apply to measures that use more than one source of data in one set of specifications/instructions (e.g., claims data to identify the denominator and medical record abstraction for the numerator).* ***Comparability is not required when comparing performance scores with and without SDS factors in the risk adjustment model. However, if comparability is not demonstrated for measures with more than one set of specifications/instructions, the different specifications (e.g., for medical records vs. claims) should be submitted as separate measures.***

2b6.1. Describe the method of testing conducted to compare performance scores for the same entities across the different data sources/specifications (*describe the steps—do not just name a method; what statistical analysis was used*)

2b6.2. What were the statistical results from testing comparability of performance scores for the same entities when using different data sources/specifications? (*e.g., correlation, rank order*)

2b6.3. What is your interpretation of the results in terms of the differences in performance measure scores for the same entities across the different data sources/specifications? (i.*e., what do the results mean and what are the norms for the test conducted*)

_____

2b7. MISSING DATA ANALYSIS AND MINIMIZING BIAS

2b7.1. Describe the method of testing conducted to identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias (*describe the steps—do not just name a method; what statistical analysis was used*) Not done

2b7.2. What is the overall frequency of missing data, the distribution of missing data across providers, and the results from testing related to missing data? (*e.g., results of sensitivity analysis of the effect of various rules for missing data/nonresponse; if no empirical sensitivity analysis, identify the approaches for handling missing data that were considered and pros and cons of each*) Over the nine years, 947 infants (0.2%) were missing a value for retinal examination. Missing data does not appear to be an issue.

2b7.3. What is your interpretation of the results in terms of demonstrating that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias**?** (i.*e., what do the results mean in terms of supporting the selected approach for missing data and what are the norms for the test conducted; if no empirical analysis, provide rationale for the selected approach for missing data*)

3. Feasibility

Extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

3a. Byproduct of Care Processes

For clinical measures, the required data elements are routinely generated and used during care delivery (e.g., blood pressure, lab test, diagnosis, medication order).

3a.1. **Data Elements Generated as Byproduct of Care Processes.**

generated by and used by healthcare personnel during the provision of care, e.g., blood pressure, lab value, medical condition, Coded by someone other than person obtaining original information (e.g., DRG, ICD-9 codes on claims), Abstracted from a record by someone other than person obtaining original information (e.g., chart abstraction for quality measure or registry)

If other:

3b. Electronic Sources

The required data elements are available in electronic health records or other electronic sources. If the required data are not in electronic health records or existing electronic sources, a credible, near-term path to electronic collection is specified.

3b.1. **To what extent are the specified data elements available electronically in defined fields?** *(i.e., data elements that are needed to compute the performance measure score are in defined, computer-readable fields)*

ALL data elements are in defined fields in a combination of electronic sources

3b.2. **If ALL the data elements needed to compute the performance measure score are not from electronic sources, specify a credible, near-term path to electronic capture, OR provide a rationale for using other than electronic sources.**

3b.3. **If this is an eMeasure, provide a summary of the feasibility assessment in an attached file or make available at a measure-specific URL.**

Attachment:

3c. Data Collection Strategy

Demonstration that the data collection strategy (e.g., source, timing, frequency, sampling, patient confidentiality, costs associated with fees/licensing of proprietary measures) can be implemented (e.g., already in operational use, or testing demonstrates that it is ready to put into operational use). For eMeasures, a feasibility assessment addresses the data elements and measure logic and demonstrates the eMeasure can be implemented or feasibility concerns can be adequately addressed.

3c.1. **Describe what you have learned/modified as a result of testing and/or operational use of the measure regarding data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, time and cost of data collection, other feasibility/implementation issues.**

IF a PRO-PM, consider implications for both individuals providing PROM data (patients, service recipients, respondents) and those whose performance is being measured.

Patient identifiers are not collected in the registry. Confidentiality for each hospital member is strictly maintained. Procedures in place assure reasonable confidence that data are complete and accurate. There are no specific fees for this measure, although members of the Vermont Oxford Network pay an annual membership fee.

3c.2. **Describe any fees, licensing, or other requirements to use any aspect of the measure as specified** *(e.g., value/code set, risk model, programming code, algorithm)*.

None

4. Usability and Use

Extent to which potential audiences (e.g., consumers, purchasers, providers, policy makers) are using or could use performance results for both accountability and performance improvement to achieve the goal of high-quality, efficient healthcare for individuals or populations.

4a. Accountability and Transparency

Performance results are used in at least one accountability application within three years after initial endorsement and are publicly reported within six years after initial endorsement (or the data on performance results are available). If not in use at the time of initial endorsement, then a credible plan for implementation within the specified timeframes is provided.

4.1. **Current and Planned Use**

NQF-endorsed measures are expected to be used in at least one accountability application within 3 years and publicly reported within 6 years of initial endorsement in addition to performance improvement.

Planned	Current Use (for current use provide URL)
Public Reporting	Quality Improvement with Benchmarking (external benchmarking to multiple organizations) Vermont Oxford Network Password protected website Quality Improvement (Internal to the specific organization) N/A N/A

4a.1. **For each CURRENT use, checked above, provide:**

- Name of program and sponsor
- Purpose
- Geographic area and number and percentage of accountable entities and patients included

Vermont Oxford Network has nearly 1000 members. We do not know what proportion are using this measure for quality improvement.

4a.2. **If not currently publicly reported OR used in at least one other accountability application (e.g., payment program, certification, licensing) what are the reasons?** (*e.g., Do policies or actions of the developer/steward or accountable entities restrict access to performance results or impede implementation?*)

There are no publicly-reported quality measures for the neonatal population. Vermont Oxford Network members use reports provided by the Network to do internal quality improvement and benchmarking on retinal exams. Additionally, Vermont Oxford Network has managed several quality improvement collaboratives for its members on reducing ROP.

4a.3. **If not currently publicly reported OR used in at least one other accountability application, provide a credible plan for implementation within the expected timeframes -- any accountability application within 3 years and publicly reported within 6 years of initial endorsement.** (*Credible plan includes the specific program, purpose, intended audience, and timeline for implementing the measure within the specified timeframes. A plan for accountability applications addresses mechanisms for data aggregation and reporting.*)

Vermont Oxford Network is committed to working with accrediting bodies that are developing publicly-reported quality measures for the neonatal population.

4b. Improvement

Progress toward achieving the goal of high-quality, efficient healthcare for individuals or populations is demonstrated. If not in use for performance improvement at the time of initial endorsement, then a credible rationale describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.

4b.1. **Progress on Improvement. (Not required for initial endorsement unless available.)**

Performance results on this measure (current and over time) should be provided in 1b.2 and 1b.4. Discuss:

- **Progress (trends in performance results, number and percentage of people receiving high-quality healthcare)**
- **Geographic area and number and percentage of accountable entities and patients included**

The proportion of infants receiving a timely retinopathy of prematurity examination has increased slightly from 2006 (90.1%) to 2014

(91.8%). The rate for hospitals at the 25th percentile increased from 88.2% in 2006 to 91.7% in 2012, decreasing to 90% in 2014.

4b.2. **If no improvement was demonstrated, what are the reasons? If not in use for performance improvement at the time of initial endorsement, provide a credible rationale that describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.**

4c. Unintended Consequences

The benefits of the performance measure in facilitating progress toward achieving high-quality, efficient healthcare for individuals or populations outweigh evidence of unintended negative consequences to individuals or populations (if such evidence exists).

4c.1. **Were any unintended negative consequences to individuals or populations identified during testing; OR has evidence of unintended negative consequences to individuals or populations been reported since implementation? If so, identify the negative unintended consequences and describe how benefits outweigh them or actions taken to mitigate them.**

Unintended negative consequences to individuals or populations are unknown. To mitigate unintended consequences associated with Vermont Oxford Network member hospitals, members receive a manual of operations annually that contains definitions and clearly operationalized criteria for the measure. Comprehensive business rules verify records for consistency, completeness and accuracy. Centers employ a definitive process to assure that the measure is not reported until data are complete and correct. Hospital contacts must verify that data for all eligible infants are submitted prior to finalization.

5. Comparison to Related or Competing Measures

If a measure meets the above criteria and there are endorsed or new related measures (either the same measure focus or the same target population) or competing measures (both the same measure focus and the same target population), the measures are compared to address harmonization and/or selection of the best measure.

5. Relation to Other NQF-endorsed Measures
Are there related measures (conceptually, either same measure focus or target population) or competing measures (conceptually both the same measure focus and same target population)? If yes, list the NQF # and title of all related and/or competing measures.
No

5.1a. **List of related or competing measures (selected from NQF-endorsed measures)**

5.1b. **If related or competing measures are not NQF endorsed please indicate measure title and steward.**

5a. Harmonization

The measure specifications are harmonized with related measures;
OR
The differences in specifications are justified

5a.1. **If this measure conceptually addresses EITHER the same measure focus OR the same target population as NQF-endorsed measure(s):**
Are the measure specifications completely harmonized?

5a.2. **If the measure specifications are not completely harmonized, identify the differences, rationale, and impact on interpretability and data collection burden.**

5b. Competing Measures

The measure is superior to competing measures (e.g., is a more valid or efficient way to measure);
OR
Multiple measures are justified.

5b.1. **If this measure conceptually addresses both the same measure focus and the same target population as NQF-endorsed measure(s):**
Describe why this measure is superior to competing measures (e.g., a more valid or efficient way to measure quality); OR provide a rationale for the additive value of endorsing an additional measure. (Provide analyses when possible.)
N/A

Appendix

A.1 Supplemental materials may be provided in an appendix. All supplemental materials (such as data collection instrument or methodology reports) should be organized in one file with a table of contents or bookmarks. If material pertains to a specific submission form number, that should be indicated. Requested information should be provided in the submission form and required attachments. There is no guarantee that supplemental materials will be reviewed.

No appendix **Attachment:**

Contact Information

Co.1 **Measure Steward (Intellectual Property Owner):** Vermont Oxford Network
Co.2 **Point of Contact:** Erika, Edwards, eedwards@vtoxford.org, 802-865-4814-246
Co.3 **Measure Developer if different from Measure Steward:** Vermont Oxford Network
Co.4 **Point of Contact:** Erika, Edwards, eedwards@vtoxford.org, 802-865-4814-246

Additional Information

Ad.1 **Workgroup/Expert Panel involved in measure development**
Provide a list of sponsoring organizations and workgroup/panel members' names and organizations. Describe the members' role in measure development.
N/A

Measure Developer/Steward Updates and Ongoing Maintenance
Ad.2 **Year the measure was first released:** 2008
Ad.3 **Month and Year of most recent revision:** 01, 2016
Ad.4 **What is your frequency for review/update of this measure?** Annual
Ad.5 **When is the next scheduled review/update for this measure?** 01, 2017

Ad.7 **Disclaimers:**

Ad.8 **Additional Information/Comments:**

MEASURE WORKSHEET

This document summarizes the evaluation of the measure as it progresses through NQF's Consensus Development Process (CDP). The information submitted by measure developers/stewards is included after the Brief Measure Information, Preliminary Analysis, and Pre-meeting Public and Member Comments sections.

To navigate the links in the worksheet: Ctrl + click link to go to the link; ALT + LEFT ARROW to return

Brief Measure Information

NQF #: 0716
Measure Title: Unexpected Complications in Term Newborns
Measure Steward: California Maternal Quality Care Collaborative
Brief Description of Measure: This is a hospital level performance score reported as the percent of infants with Unexpected Newborn Complications among full term newborns with no preexisting conditions, typically calculated per year.
Developer Rationale: The most important childbirth outcome for families is bringing home a healthy baby. While there have been measures developed to assess clinical practices and outcomes in preterm infants, there are a lack of metrics that assess the health outcomes of term infants who represent over 90% of all births.
The Unexpected Complications in Term Newborns metric addresses this gap and measures adverse outcomes resulting in severe or moderate morbidity in otherwise healthy term infants without preexisting conditions. Importantly, this metric also serves as a balancing measure for other NQF endorsed maternal measures such as NTSV Cesarean rates, third and fourth degree lacerations, episiotomies and early elective delivery rates. The purpose of a balancing measure to is guard against any unanticipated or unintended consequences of quality improvement activities for these measures.

Numerator Statement: Numerator: The numerator is divided into two categories: Severe complications and moderate complications.
Severe complications include neonatal death, transfer to another hospital for higher level of care, extremely low Apgar Scores (=3 at either 5 or 10 minutes of life), severe birth injuries such as intracranial hemorrhage or nerve injury, neurologic damage, severe respiratory and infectious complications such as sepsis. Parents of such babies may often worry about short or long term infant outcomes.
Moderate complications include diagnoses or procedures that raise concern but at a lower level than the list for severe (e.g. use of CPAP or bone fracture). For inclusion in the numerator, most require an infant length of stay that exceeds that of the mother, validating that these are indeed significant complications. Examples include less severe respiratory complications (e.g. Transient Tachypnea of the Newborn), or infections with a longer length of stay not including sepsis.
As a "safety net" to capture cases who were under-coded, the numerator also includes infants who have a prolonged length of stay of over 5 days to capture the "seemingly normal" infants with neither any form of jaundice nor a social reason for staying in the hospital (e.g. family disruption or adoption).
Denominator Statement: The denominator is comprised of singleton, live born babies who are at least 37.0 weeks of gestation, and over 2500g in birth weight. The denominator excludes most serious fetal conditions that are "preexisting" (present before labor), including prematurity, multiple gestations, poor fetal growth, congenital malformations, genetic disorders, other specified fetal and maternal conditions and infants exposed to maternal drug use in-utero. The final denominator population consists of babies who are expected to do well following labor and delivery and go home routinely with their mothers.
Denominator Exclusions: a) Babies not born in hospitals are excluded as this is a hospital quality performance measure
b) Babies who are part of multiple gestation pregnancies are excluded.
c) Premature infants (babies born before 37 weeks gestational age) are excluded
d) Low birth weight babies (<=2500g) are excluded
e) Babies with congenital malformations and genetic diseases are excluded
f) Babies with pre-existing fetal conditions such as IUGR are excluded

g) Babies who were exposed to maternal drug use in-utero are excluded

Measure Type: Outcome
Data Source: Administrative claims
Level of Analysis: Facility, Integrated Delivery System, Population : Regional, Population : State

IF Endorsement Maintenance – Original Endorsement Date: Jan 17, 2011 **Most Recent Endorsement Date:** Jan 17, 2011

Maintenance of Endorsement -- Preliminary Analysis

This measure was originally titled "Healthy Term Newborn"

To maintain NQF endorsement endorsed measures are evaluated periodically to ensure that the measures still meets the NQF endorsement criteria ("maintenance"). The emphasis for maintaining endorsement is focused on how effective the measure is for promoting improvements in quality. Endorsed measures should have some experience from the field to inform the evaluation. The emphasis for maintaining endorsement is noted for each criterion.

Criteria 1: Importance to Measure and Report

1a. Evidence
Maintenance measures – less emphasis on evidence unless there is new information or change in evidence since the prior evaluation.

1a. Evidence. The evidence requirements for a health outcomes measure include providing rationale that supports the relationship of the health outcome to processes or structures of care. The guidance for evaluating the clinical evidence asks if the relationship between the measured health outcome and at least one clinical action is identified and supported by the stated rationale.

Summary of evidence:

- The developer provides a diagram of the linkage between the health outcome and the healthcare processes that influence it.
- They provide evidence for two examples of health care interventions during labor that can have an impact on the health outcome (a healthy term newborn): screening during labor for Group B strep and provision of prophylaxis, and proper use of vacuum extractors, both supported by clinical practice guidelines.

Changes to evidence from last review

☐ **The developer attests that there have been no changes in the evidence since the measure was last evaluated.**

☒ **The developer provided updated evidence for this measure.**

Question for the Committee:

- *Although the guidelines have been updated/new studies have been provided, the underlying evidence presented appears to be the directionally the same/stronger since the last NQF endorsement review. Does the Committee agree and so there is no need for repeat discussion and vote on Evidence?*

Preliminary rating for evidence: ☒ **Pass** ☐ **No Pass**

1b. Gap in Care/Opportunity for Improvement and 1b. Disparities
Maintenance measures – increased emphasis on gap and variation

1b. Performance Gap. The performance gap requirements include demonstrating quality problems and opportunity for improvement.

- The developer states that "While there have been measures developed to assess clinical practices and outcomes in preterm infants, there are a lack of metrics that assess the health outcomes of term infants who represent over 90% of all births." They also note this measure is intended to serve as a balancing measure for the other

childbirth-related measures.
Data from California:

Year, # hospitals, and # births	Mean (standard deviation)	Min-Max	Interquartile range
2011: 226, 389,741	3.56 (2.14)	0.68-15.19	25th percentile: 2.16 75th percentile: 4.39
2012: 232, 395,634	3.63 (2.14)	0.97-14.33	25th percentile: 2.18 75th percentile: 4.72

Disparities At the 2011 initial endorsement evaluation the Committee noted that the measure does not account for disadvantaged populations according to race, socioeconomic status, or living conditions and suggested that future testing based on stratification be conducted.

Data from California:

	2011 measure results	**2012 measure results**
#hospitals/#newborns	**247/392,328**	**?/397,864**
Male	**3.96%**	**4.04%**
Female	**2.99%**	**3.07%**
Hispanic, native born	**3.39%**	**3.53%**
Hispanic. Foreign born	**3.23%**	**3.36%**
Non-hispanic white	**3.52%**	**3.60%**
Non-hispanic black	**4.25%**	**4.26%**
Asian/Pacific Islander	**3.35%**	**3.28%**
Other race/ethnicity	**5.31%**	**5.50%**
Urban	**3.48%**	**3.58%**
Rural	**3.47%**	**3.46%**
Medicaid	**3.51%**	**3.62%**
Private	**3.28%**	**3.38%**
Self-pay	**2,43%**	**2.18%**
Other payment	**8.98%**	**8.14%**
Hospital – city/county	**3.90%**	**4.37%**
Hospital - For profit	**2.94%**	**2.76%**
Hospital – Not for Profit/University	**3.51%**	**3.57%**
Hospital – unknown ownership	**3.55%**	**3.28%**
Basic/No NICU	**2.94%**	**2,81%**
Intermediate NICU	**3.45%**	**3.51%**
Community NICU	**3.54%**	**3.61%**
Regional NICU	**4.53%**	**5.05%**

Questions for the Committee:

- *Is data from 2011 and 2012 sufficient to understand current performance?*
- *Is there a gap in care that warrants a national performance measure?*
- *Can this measure be used to understand disparities in perinatal care?*

Preliminary rating for opportunity for improvement: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 1: Importance to Measure and Report (including 1a, 1b, 1c)

1a. Evidence to Support Measure Focus

Comments:

Agree that there is sufficient evidence to support the measure.

1b. Performance Gap

Comments:

It might be helpful to have more recent data, given that the outcomes are often directly related to the intrapartum period. Since this is a balancing measure, it would held to see if the push for improved safety on L&D is working... This is a measure that would definite help to elucidate disparities.

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability

2a1. Reliability Specifications

Maintenance measures – no change in emphasis – specifications should be evaluated the same as with new measures

2a1. Specifications requires the measure, as specified, to produce consistent (reliable) and credible (valid) results about the quality of care when implemented.

Data source(s): Administrative claims and birth certificate data

Specifications: **The specifications have been inverted from the original endorsed version.**

- The numerator is divided into two categories: Severe complications and moderate complications.
  - Severe complications include neonatal death, transfer to another hospital for higher level of care, extremely low Apgar Scores (=3 at either 5 or 10 minutes of life), severe birth injuries such as intracranial hemorrhage or nerve injury, neurologic damage, severe respiratory and infectious complications such as sepsis. Parents of such babies may often worry about short or long term infant outcomes.
  - Moderate complications include diagnoses or procedures that raise concern but at a lower level than the list for severe (e.g. use of CPAP or bone fracture). For inclusion in the numerator, most require an infant length of stay that exceeds that of the mother, validating that these are indeed significant complications.
- The denominator is comprised of singleton, live born babies who are at least 37.0 weeks of gestation, and over 2500g in birth weight. The denominator excludes most serious fetal conditions that are "preexisting" (present before labor), including prematurity, multiple gestations, poor fetal growth, congenital malformations, genetic disorders, other specified fetal and maternal conditions and infants exposed to maternal drug use in-utero. The final denominator population consists of babies who are expected to do well following labor and delivery and go home routinely with their mothers.
- There are seven categories of exclusions:
  a) Babies not born in hospitals are excluded as this is a hospital quality performance measure
  b) Babies who are part of multiple gestation pregnancies are excluded.
  c) Premature infants (babies born before 37 weeks gestational age) are excluded
  d) Low birth weight babies (<=2500g) are excluded
  e) Babies with congenital malformations and genetic diseases are excluded
  f) Babies with pre-existing fetal conditions such as IUGR are excluded
  g) Babies who were exposed to maternal drug use in-utero are excluded
- Calculation algorithms are included.
- A number of changes were made to the specifications since the last review in 2010, including a change to the name,

- adding two sub measures (severe and moderate),
- identification of term infants - Available ICD-9 codes for gestational age and birth weight are significantly under-utilized in practice. However, the use of a linked PDD to Birth Certificate data files has solved this issue as the Birth Certificate data fields for Best Obstetric Gestational Age and Birth weight have high degrees of completeness and accuracy.
- additional information on codes such as more codes for exclusions and checks and balances to allow for differing code practices.

- ICD-10 codes are included and the measure has been updated.
- The measure is not risk adjusted.
- The measure does not require sampling, but the developer recommends that hospitals have at least 200 cases in the denominator population.

Questions for the Committee:

- *Are the changes to the specifications appropriate?*
- *Are all the data elements clearly defined? Are all appropriate codes included?*
- *Is the logic or calculation algorithm clear?*
- *Is it likely this measure is consistently implemented?*

2a2. Reliability Testing Testing attachment
Maintenance measures – less emphasis if no new testing data provided

2a2. Reliability testing demonstrates if the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise enough to distinguish differences in performance across providers.

For maintenance measures, summarize the reliability testing from the prior review:

Describe any updates to testing - new empirical reliability testing is described in RED in the attachment

SUMMARY OF TESTING

Reliability testing level ☒ **Measure score** ☐ **Data element** ☐ **Both**

Reliability testing performed with the data source and level of analysis indicated for this measure ☒ **Yes** ☐ **No**

Method(s) of reliability testing

- Signal-to-noise testing was completed using a beta-binomial model. This is an appropriate metric for hospital-level testing.
- Reliability testing was completed using data on 387,677 babies at the 226 hospitals in California that had more than 200 recorded live births that met the denominator inclusion criteria in 2011. A summary of the characteristics of the hospitals and the babies are included.

Results of reliability testing

Reliability statistics by hospital decile and by each hospital were provided.
The mean reliability score for the 226 hospitals was 0.91. Generally, reliability estimates >0.70 are considered acceptable.

Guidance from the Reliability Algorithm

Precise specifications (Box 1) → Empirical reliability testing (Box 2) → Measure score testing (Box 5) → appropriate method- STN (Box 6) → high or moderate confidence of reliability of numerator data element → rate as high

Questions for the Committee:

○ *Is the test sample adequate to generalize for widespread implementation?*

○ *Do the results demonstrate sufficient reliability so that differences in performance can be identified?*

Preliminary rating for reliability: ☒ **High** ☐ **Moderate** ☐ **Low** ☐ **Insufficient**

2b. Validity
Maintenance measures – less emphasis if no new testing data provided

2b1. Validity: Specifications

2b1. Validity Specifications. This section should determine if the measure specifications are consistent with the evidence.

Specifications consistent with evidence in 1a. ☒ **Yes** ☐ **Somewhat** ☐ **No**
Specification not completely consistent with evidence

Question for the Committee:

○ *Are the specifications consistent with the evidence?*

2b2. Validity testing

2b2. Validity Testing should demonstrate the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality.

For maintenance measures, summarize the validity testing from the prior review: face validity only.

Describe any updates to validity testing new empirical and face validity testing is described in RED in the attachment

SUMMARY OF TESTING

Validity testing level ☒ **Measure score** ☐ **Data element testing against a gold standard** ☐ **Both**

Method of validity testing of the measure score:

☒ **Face validity only**
☒ **Empirical validity testing of the measure score**

Validity testing method:

- **Empirical validity** testing was conducted by the adaptive Neyman test, and Pearson and Spearman correlations were used to demonstrate the relationship between the Unexpected Newborn Complications measure and admission to the NICU. The developer states "Neonatal admission to the NICU is a very good indication of an unexpected newborn complication in a term infant with no preexisting conditions." NPIC uses the Unexpected Newborn Complications metric for performance and benchmarking in their member hospitals. One of the variables present in the NPIC dataset is "admission to special care nursery". This allows NPIC to compare the unexpected newborn complications measure to admissions to special care nursery in their member hospitals.
- **Empirical validity** testing of the measure as a balancing measure for three QI projects that reduced the NTSV Cesarean birth rate.
- **Face validity** was assessed by a group of 22 experts who were not involved in developing the measure.

Validity testing results:

- The results from applying the adaptive Neyman test to the unexpected newborn complications odds ratios (stratified by gestational age) to a very similar measure (admissions to the neonatal intensive care unit in babies) in the NPIC cohort of babies demonstrate that the means of the odds ratios are very similar.
- In three hospital QI projects that reduced NTSV CS rates, the Unexpected Complications in Term Newborns (*UNC) rate also declined*, reassuring the medical staff that the intervention on the mother did not produce an adverse outcome in the baby. This demonstrates the utility of the UNC measure as a balancing measure for NTSV Cesarean section.

- The face validity panel was asked to assess the measure on a 1-5 scale. The mean rating was 4.82/5, and 19/22 rated it a 5, indicating a high level of face validity.

Questions for the Committee:

- *Is the test sample adequate to generalize for widespread implementation?*
- *Do the results demonstrate sufficient validity so that conclusions about quality can be made?*
- *Do you agree that the score from this measure as specified is an indicator of quality?*

2b3-2b7. Threats to Validity

2b3. Exclusions:

The developers analyzed the following exclusions for frequency and variability across hospitals:

- Low Birth Weight Babies (under 2500g)
- Preterm births (under 37) weeks and births above 47 weeks of gestation (to exclude nonsensical values of gestational age).
- Babies with Congenital Malformations (list of ICD-9 codes outlined in Appendix 2A)
- Babies with Pre-existing Conditions (list of ICD-9 codes outlined in Appendix 2B)
- Babies exposed to maternal drug use in-utero (list of ICD-9 codes outlined in Appendix 2C)

They include tables for 2011 and 2012 data. Overall frequency is 12.8% for 2011 and 13.1% for 2012. The developers state the distribution is as expected and random.

Rationales for each exclusion are included.

Questions for the Committee:

- *Are the exclusions consistent with the evidence?*
- *Are any patients or patient groups inappropriately excluded from the measure? Are there additional patients that should be excluded?*
- *Are the exclusions/exceptions of sufficient frequency and variation across providers to be needed (and outweigh the data collection burden)?*

2b4. Risk adjustment: **Risk-adjustment method** ☒ **None** ☐ **Statistical model** ☐ **Stratification**

- The developers provide analysis of potential risk adjustment but conclude "In summary, while there are some individual factors that can statistically affect the score, when examined together at the hospital level they cancel each other out or are distributed evenly among hospitals so as not to significantly affect the rankings."

Questions for the Committee:

- *Is the developer's rationale that there is no conceptual basis for adjusting this measure for SDS factors acceptable?*

2b5. Meaningful difference *(can statistically significant and clinically/practically meaningful differences in performance measure scores can be identified)*:

- Hospitals were identified as statistically significantly better or worse than the state average had scores that were at least 20% lower or 20% higher than the state mean, which the developers consider a meaningful difference in performance.
- 53 hospitals (21.5%) of 247 California hospitals were rated as statistically significantly higher (worse) than the state mean and 78 hospitals (31.5%) were identified as statistically significantly lower (better) than the state mean.
- Summary statistics for all 247 hospitals:

Mean	SD	Minimum	10th percentile	25th percentile	Median	75th percentile	90th percentile	Maximum
3.48	1.87	0.0	1.59	2.18	3.13	4.29	6.04	15.2

Question for the Committee:

- *Does this measure identify meaningful differences about quality?*

2b6. Comparability of data sources/methods: N/A

2b7. Missing Data

- Four variables are essential to calculate the measure: codes to identify in-hospital singleton births; missing gestational age; missing birth weight; under-coding of diagnoses.
- Methods for finding this information in other ways are outlined to minimize the number of missing cases.
- The developer concludes that "Missing data elements are too miniscule to have any effect on the measure. As stated in our responses to questions 2b7.1 and 2b7.2, the unexpected newborn complications measure has inbuilt checks to account for possible missing data, as well as under and over coding. It is highly unlikely that our measure will miss including an eligible infant."

Guidance from the algorithm:
Specifications aligned with evidence (Box 1) → Threats to validity assessed (Box 2) →empirical validity testing (Box 3) → Testing of measure score (Box 6) → appropriate method (Box 7) →moderate or high certainty (Box 8a 8b) → moderate or high

Preliminary rating for validity: ☒ **High** ☐ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 2: Scientific Acceptability of Measure Properties (including all 2a, 2b, and 2d)

2a1. & 2b1. Specifications
Comments:
The sub measures are helpful as is updating with ICD 10. Adding the Birth Certificate data helps with consistently implementation.
I agree with the score.

2a2. Reliability Testing
Comments:
I agree with the score.

2b2. Validity Testing
Comments:
See comments

2b3. Exclusions Analysis
2b4. Risk Adjustment/Stratification for Outcome or Resource Use Measures
2b5. Identification of Statistically Significant & Meaningful Differences In Performance
2b6. Comparability of Performance Scores When More Than One Set of Specifications
2b7. Missing Data Analysis and Minimizing Bias
Comments:
Adding data from EMR may enhance the validity. Often what is billed (claims) and what is documented is not always congruent. Since EMR is not yet universal and, there are too many different platforms, adding this to the data collection will make the measure less feasible.

Criterion 3. Feasibility
Maintenance measures – no change in emphasis – implementation issues may be more prominent

3. Feasibility is the extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

- The measure originally used Patient Discharge Data only, but it was found that it was too difficult to identify term infants. However, linking PDD to Birth Certificate data files has solved this issue as the Birth Certificate data fields for Best Obstetric Gestational Age and Birth weight have high degrees of completeness and accuracy. The Birth Certificate also allows the addition of extremely low Apgar scores as an additional numerator item.
- The developer also learned that coding practices vary, with some overcoding and others undercoding. They have included additional codes to account for this.

Questions for the Committee:

- *Are the required data elements routinely generated and used during care delivery?*
- *Are the required data elements available in electronic form, e.g., EHR or other electronic sources?*
- *Is the data collection strategy ready to be put into operational use?*
- *Have the developers adequately addressed the identified issues?*

Preliminary rating for feasibility: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 3: Feasibility

3a. Byproduct of Care Processes
3b. Electronic Sources
3c. Data Collection Strategy
Comments:
Claims data is easy to access and as such is feasible.

Criterion 4: Usability and Use
Maintenance measures – increased emphasis – much greater focus on measure use and usefulness, including both impact /improvement and unintended consequences

4. Usability and Use evaluate the extent to which audiences (e.g., consumers, purchasers, providers, policymakers) use or could use performance results for both accountability and performance improvement activities.

Current uses of the measure
Publicly reported? ☐ **Yes** ☒ **No**

Current use in an accountability program? ☐ **Yes** ☐ **No**
OR
Planned use in an accountability program? ☒ **Yes** ☐ **No**

Accountability program details

- The measure is currently in use for internal quality improvement and internal QI with benchmarking.
- The developers state that "This measure was being tweaked in response to user feedback and we now feel that it is ready for public reporting. Our philosophy has been to provide hospitals with their own data for 2 to 3 years and then begin public reporting of hospital results."
- Public reporting is planned to begin this year in California, with 2015 data.
- The measure specifications are available free of cost on a public website.

Improvement results

- The developers state they have been revising and improving the measure based on user feedback. In 2011, it was only used in California, but has now been used in Washington, Oregon, Alaska, and Montana hospitals, as well as National Perinatal Information Center (NPIC)'s 82 member hospitals across the country. It is now

evaluating 1 million births across the US (25%).

- State wide improvement results are not yet available, however, they have seen individual improvement at hospitals who have used the measure internally (10-20% reduction) and they note that it has been useful as a balancing measure with cesarean section reduction efforts.
- The developers note that as planned, this "Allowed hospitals 2-3 years of internal data benchmarking before we provide hospital level results for public reporting to the California Hospital Accountability and Reporting Taskforce (CHART), supported by the California HealthCare Foundation."

Unexpected findings (positive or negative) during implementation
No intended findings or potential harms have been noted.

Potential harms
The developer is "encouraging that this metric be used as a balancing measure to identify unintended consequences of other measures such as Cesarean Sections."

Feedback:

- In 2013 MAP reviewed this measure noting some technical concerns needed more experience with the measure but strongly supported the direction of the measure for use in federal programs.

***Questions for the Committee**:*

- *Has this measure been well vetted by those being measured and other prior to public reporting?*
- *How can the performance results be used to further the goal of high-quality, efficient healthcare?*
- *Do the benefits of the measure outweigh any potential unintended consequences?*

Preliminary rating for usability and use: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 4: Usability and Use

4a. Accountability and Transparency
4b. Improvement
4c. Unintended Consequences
Comments:

Criterion 5: Related and Competing Measures

Related or competing measures
There are no related/competing measures.

Harmonization
N/A

Pre-meeting public and member comments

-

Measure Title: Click here to enter measure title

IF the measure is a component in a composite performance measure, provide the title of the Composite Measure here: Click here to enter composite measure title

Date of Submission: Click here to enter a date

Instructions

- *For composite performance measures:*
  - *A separate evidence form is required for each component measure unless several components were studied together.*
  - *If a component measure is submitted as an individual performance measure, attach the evidence form to the individual measure submission.*
- Respond to all questions as instructed with answers immediately following the question. All information needed to demonstrate meeting the evidence subcriterion (1a) must be in this form. An appendix of *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 10 pages (*incudes questions/instructions*; minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.

Note: The information provided in this form is intended to aid the Steering Committee and other stakeholders in understanding to what degree the evidence for this measure meets NQF's evaluation criteria.

Subcriterion 1a. Evidence to Support the Measure Focus

The measure focus is a health outcome or is evidence-based, demonstrated as follows:

- Health outcome:[3] a rationale supports the relationship of the health outcome to processes or structures of care.
- Intermediate clinical outcome, Process,[4] or Structure: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence[5] that the measure focus leads to a desired health outcome.
- Patient experience with care: evidence that the measured aspects of care are those valued by patients and for which the patient is the best and/or only source of information OR that patient experience with care is correlated with desired outcomes.
- Efficiency:[6] evidence for the quality component as noted above.

Notes

3. Generally, rare event outcomes do not provide adequate information for improvement or discrimination; however, serious reportable events that are compared to zero are appropriate outcomes for public reporting and quality improvement.

4. Clinical care processes typically include multiple steps: assess → identify problem/potential problem → choose/plan intervention (with patient input) → provide intervention → evaluate impact on health status. If the measure focus is one step in such a multistep process, the step with the strongest evidence for the link to the desired outcome should be selected as the focus of measurement.

5. The preferred systems for grading the evidence are the U.S. Preventive Services Task Force (USPSTF) grading definitions and methods, or Grading of Recommendations, Assessment, Development and Evaluation (GRADE) guidelines.

6. Measures of efficiency combine the concepts of resource use and quality (NQF's Measurement Framework: Evaluating Efficiency Across Episodes of Care; AQA Principles of Efficiency Measures).

1a.1.This is a measure of:

Outcome

☐ Health outcome: Click here to name the health outcome

Health outcome includes patient-reported outcomes (PRO, i.e., HRQoL/functional status, symptom/burden, experience with care, health-related behaviors)

☐ Intermediate clinical outcome: Click here to name the intermediate outcome

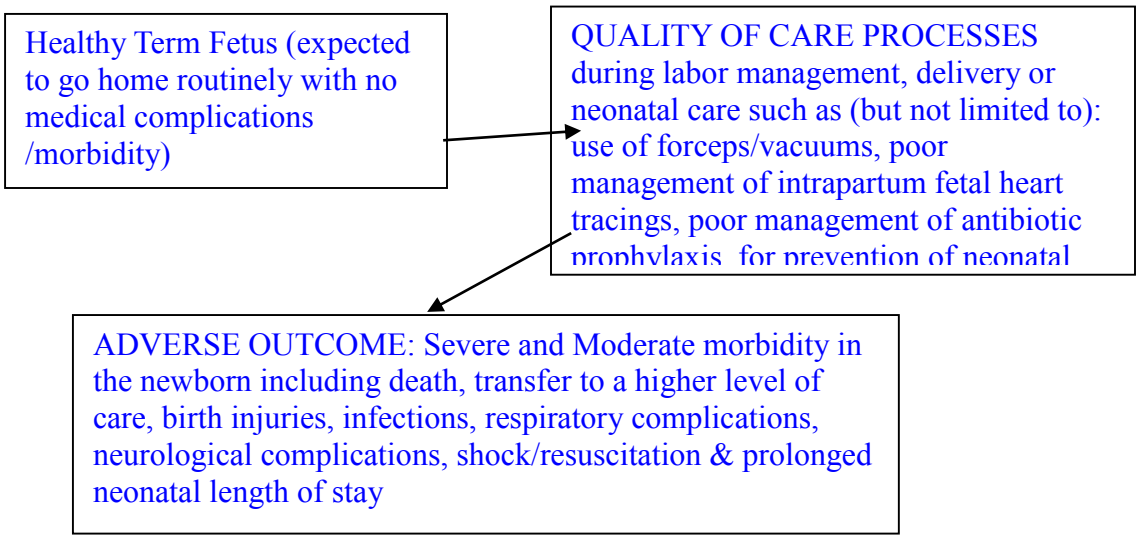
☐ Process: Click here to name the process
☐ Structure: Click here to name the structure
☐ Other: Quality of obstetric and neonatal care at the hospital level

HEALTH OUTCOME PERFORMANCE MEASURE *If not a health outcome, skip to 1a.3*

1a.2. Briefly state or diagram the linkage between the health outcome (or PRO) and the healthcare structures, processes, interventions, or services that influence it.

In the panel of maternal and child health outcomes, this measure is the one that evaluates the largest population of live births; healthy term newborns. This metric also serves as a balancing measure for other maternal and neonatal procedures and measures (i.e.) if you increase (or decrease) one does that affect baby outcomes? For example, many modern obstetric practices (such as the use of inductions, vacuums, forceps and cesarean deliveries) are done in the name of improving baby outcomes without having a proper measure to document that. In fact many of these interventions when formally studied actually lead to a diminution of newborn health.

Multiple care processes can influence rapid deterioration in a newborn's health status during labor management, delivery or neonatal care resulting in unexpected severe or moderate morbidity for the newborn with short or long term consequences.



1a.2.1. State the rationale supporting the relationship between the health outcome (or PRO) and at least one healthcare structure, process, intervention, or service.

We have outlined two examples of health care interventions during labor that can have an impact on the health outcome (a healthy term newborn).

1. Screening for maternal vaginal colonization of Group B Streptococcus (GBS) and antibiotic prophylaxis during labor will reduce neonatal infections.
2. Proper use of vacuum extractors respecting the time duration and the number of pop-offs reduces the incidence of birth trauma.

Note: For health outcome performance measures, no further information is required; however, you may provide evidence for any of the structures, processes, interventions, or service identified above.

INTERMEDIATE OUTCOME, PROCESS, OR STRUCTURE PERFORMANCE MEASURE

1a.3. Briefly state or diagram the linkages between structure, process, intermediate outcome, and health outcomes. Include all the steps between the measure focus and the health outcome.

N/A

1a.3.1. What is the source of the systematic review of the body of evidence that supports the performance measure?
☐ Clinical Practice Guideline recommendation – ***complete sections 1a.4, and 1a.7***
☐ US Preventive Services Task Force Recommendation – ***complete sections 1a.5 and 1a.7***
☐ Other systematic review and grading of the body of evidence (*e.g., Cochrane Collaboration, AHRQ Evidence Practice Center*) – ***complete sections 1a.6 and 1a.7***
☐ Other – ***complete section 1a.8***

Please complete the sections indicated above for the source of evidence. You may skip the sections that do not apply.

1a.4. CLINICAL PRACTICE GUIDELINE RECOMMENDATION
1a.4.1. Guideline citation (*including date*) and **URL for guideline** (*if available online*):

1) Shoulder dystocia. ACOG Practice Bulletin No. 40. American College of Obstetrician and Gynecologists. Obstet Gynecol 2002;100:1045-50
http://www.acog.org/~/media/Practice%20Bulletins/Committee%20on%20Practice%20Bulletins%20--%20Obstetrics/pb040.pdf?dmc=1&ts=20130906T1432320924

2) Induction of Labor. ACOG Practice Bulletin No. 107. American College of Obstetricians and Gynecologists. Obstet Gynecol 2009;114:386-97
http://www.acog.org/Resources_And_Publications/Practice_Bulletins/Committee_on_Practice_Bulletins_--_Obstetrics/Induction_of_Labor

3) Management of intrapartum fetal heart tracings. Practice Bulletin No.116. American College of Obstetricians and Gynecologists. Obstet Gynecol 2010;116:1232-40
http://www.acog.org/Resources_And_Publications/Practice_Bulletins/Committee_on_Practice_Bulletins_--_Obstetrics/Management_of_Intrapartum_Fetal_Heart_Rate_Tracings

4) Operative Vaginal Delivery. Practice Bulletin No. 17. American College of Obstetricians and Gynecologists. Obstet Gynecol 2000;
http://www.acog.org/Resources_And_Publications/Practice_Bulletins/Committee_on_Practice_Bulletins_--_Obstetrics/Operative_Vaginal_Delivery

5) Centers for Disease Control and Prevention. Prevention of Perinatal Group B Streptococcal Disease. MMWR 2010;59 (No. RR 10): 1-36
http://www.cdc.gov/mmwr/pdf/rr/rr5910.pdf

1a.4.2. Identify guideline recommendation number and/or page number and **quote verbatim, the specific guideline recommendation**.
There are a number of obstetric guidelines that direct elements of care that in turn affect the newborn. A few examples are listed below namely,

Shoulder Dystocia (pg 3):

- "Elective induction of labor or elective cesarean delivery for all women suspected of carrying a fetus with macrosomia is not appropriate" (Level B)
- "In patients with a history of shoulder dystocia, estimated fetal weight, gestational age, maternal glucose intolerance, and the severity of the prior neonatal injury should be evaluated and the risks and benefits of cesarean delivery discussed with the patient" (Level C)

- "Planned cesarean delivery to prevent shoulder dystocia may be considered for suspected fetal macrosomia with estimated fetal weights exceeding 5,000g in women without diabetes and 4,500g in women with diabetes". (Level C)

Induction of Labor (pg 8):

- The use of misoprostol in women with prior cesarean delivery or major uterine surgery has been associated with an increase in uterine rupture and, therefore, should be avoided in the third trimester. (Level A)

Management of Intrapartum Fetal Heart Rate Tracings: pg 7

- A Category III FHR Tracing is abnormal and conveys an increased risk of fetal academia at the time of observation (Level A).
- Category II tracings require evaluation, continued surveillance, initiation of appropriate corrective measures when indicated, and reevaluation. The presence of FHR accelerations (whether spontaneous or elicited) or moderate FHR variability or both are highly predictive of normal fetal acid-base status and, thus, may help guide clinical management. (Level B)

Operative Vaginal Delivery Recommendations: pg 6-7

- "The vacuum extractor is associated with an increased incidence of neonatal cephalohematomata, retinal hemorrhages, and jaundice when compared with forceps delivery" (Level A)
- "Operators should attempt to minimize the duration of vacuum application, because cephalohematomata is more likely to occur as the interval increases: (Level B)
- "The incidence of intracranial hemorrhage is highest among infants delivered by cesarean following a failed vacuum or forceps delivery. The combination of vacuum and forceps has a similar incidence of intracranial hemorrhage. Therefore, an operative vaginal delivery should not be attempted when the probability of success is very low" (Level B)

Prevention of Perinatal Group B Streptococcal Disease, Revised Guidelines from CDC 2010
Intrapartum Antibiotic Prophylaxis: (pg 17, 21)

- "Intrapartum antibiotic prophylaxis agents and dosing should be administered according to the recommendations provided in Figure 8."

Prevention of Perinatal Group B Streptococcal Disease, Revised Guidelines from CDC 2010
Secondary Prevention of Potential Sepsis in Newborns: (pg 21, 22)

- "To detect potential sepsis cases in newborns as early as possible, newborns should be managed according to the algorithm provided (Figure 9). The following are key components of the neonatal management algorithm:"
- "Any newborn with signs of sepsis should receive a full diagnostic evaluation and receive antibiotic therapy pending the results of the evaluation. The evaluation should include a blood culture; a CBC including white blood cell differential and platelet count; a chest radiograph if any abnormal respiratory signs are present; and a lumbar puncture if the newborn is stable enough to tolerate the procedure and sepsis is suspected. Therapy for the infant should include antimicrobial agents active against GBS (including intravenous ampicillin) as well as other organisms that might cause neonatal sepsis, such as E. coli (AII)."
- "Well-appearing newborns whose mothers had suspected chorioamnionitis should undergo a limited evaluation and receive antibiotic therapy pending culture results (AII). The evaluation should include a blood culture and a CBC including white blood cell differential and platelet count; no chest radiograph or lumbar puncture is needed. Consultation with obstetric providers to assess whether chorioamnionitis was suspected is important to determine neonatal management (CIII)."
- "Well-appearing infants whose mothers had no chorioamnionitis and no indication for GBS prophylaxis should be managed according to routine clinical care (CIII)."
- "Well-appearing infants of any gestational age whose mother received adequate intrapartum GBS prophylaxis (≥4 hours of penicillin, ampicillin, or cefazolin before delivery) should be observed for ≥48 hours, and no routine

diagnostic testing is recommended (BIII). Such infants can be discharged home as early as 24 hours after delivery, assuming that other discharge criteria have been met, ready access to medical care exists, and that a person able to comply fully with instructions for home observation will be present (CIII)."

- "For well-appearing infants born to mothers who had an indication for GBS prophylaxis but received no or inadequate prophylaxis, if the infant is well-appearing and ≥37 weeks and 0 days' gestational age and the duration of membrane rupture before delivery was <18 hours, then the infant should be observed for ≥48 hours, and no routine diagnostic testing is recommended (BIII). If the infant is well-appearing and either <37 weeks and 0 days' gestational age or the duration of membrane rupture before delivery was ≥18 hours, then the infant should undergo a limited evaluation and observation for ≥48 hours (BIII)."

1a.4.3. Grade assigned to the quoted recommendation with definition of the grade:

Grades are listed beside the quoted recommendations above. Definitions of the grades are provided below.

Definitions used by the American Congress of Obstetricians and Gynecologists are as follows:
Level A: Recommendations are based on good and consistent Scientific Evidence.
Level B: Recommendations are based on limited or inconsistent Scientific Evidence.
Level C: Recommendations are based primarily on consensus and expert opinion.

Definitions used by the Centers for Disease Control and Prevention

CATEGORY	DEFINITION	RECOMMENDATION
Category A:	Strong Evidence for efficacy and substantial clinical benefit	Strongly recommended
Category B:	Strong or moderate evidence for efficacy but only limited Clinical benefit.	Generally Recommended
Category C:	Insufficient evidence for efficacy or efficacy does not Outweigh possible adverse consequences.	Optional
Category D:	Moderate evidence against efficacy or for adverse outcome	Generally not recommended
Category E:	Strong evidence against efficacy or for adverse outcome	Never recommended

1a.4.4. Provide all other grades and associated definitions for recommendations in the grading system. (*Note: If separate grades for the strength of the evidence, report them in section 1a.7.*)

Grades for the Strength of Evidence are reported in section 1a.7.

1a.4.5. Citation and URL for methodology for grading recommendations (*if different from 1a.4.1*)**:**

See 1a.4.1

1a.4.6. If guideline is evidence-based (rather than expert opinion), are the details of the quantity, quality, and consistency of the body of evidence available (e.g., evidence tables)?

☐ Yes → ***complete section 1a.7***

☐ No → ***report on another systematic review of the evidence in sections 1a.6 and 1a.7; if another review does not exist, provide what is known from the guideline review of evidence in 1a.7***

1a.5. UNITED STATES PREVENTIVE SERVICES TASK FORCE RECOMMENDATION

1a.5.1. Recommendation citation (*including date*) and **URL for recommendation** (*if available online*):
n/a

1a.5.2. Identify recommendation number and/or page number and **quote verbatim, the specific recommendation**.
Not applicable

1a.5.3. Grade assigned to the quoted recommendation with definition of the grade:
N/a
1a.5.4. Provide all other grades and associated definitions for recommendations in the grading system. (*Note: the grading system for the evidence should be reported in section 1a.7.*)
n/a
1a.5.5. Citation and URL for methodology for grading recommendations (*if different from 1a.5.1*)**:**
n/a
Complete section 1a.7

1a.6. OTHER SYSTEMATIC REVIEW OF THE BODY OF EVIDENCE
1a.6.1. Citation (*including date*) and **URL** (*if available online*):
n/a

1a.6.2. Citation and URL for methodology for evidence review and grading (*if different from 1a.6.1*)**:**
n/a
Complete section 1a.7

1a.7. FINDINGS FROM SYSTEMATIC REVIEW OF BODY OF THE EVIDENCE SUPPORTING THE MEASURE
1a.7.1. What was the specific structure, treatment, intervention, service, or intermediate outcome addressed in the evidence review?
The following obstetric and neonatal treatments/interventions were addressed in each of the clinical practice guidelines and the CDC guidelines listed in Section **1a.4.1.** These practices (and others) are known to effect neonatal outcomes if not managed appropriately during labor, delivery and after birth.

1) Recommendations to reduce/prevent shoulder dystocia (a birth injury) during delivery of the newborn.
2) Recommendations on how to appropriately use vacuums to reduce the incidence of cephalohematomata, retinal hemorrhages, and jaundice in the newborn. Recommendations on how to use forceps to reduce the incidence of intracranial hemorrhage in newborns.
3) Recommendations on how to appropriately manage intrapartum fetal heart tracings specifically Category III FHR Tracing abnormal tracings which indicate fetal acidosis.
4) Specific algorithms to identify and reduce potential neonatal sepsis firstly through intrapartum antibiotic prophylaxis and also secondary prevention of sepsis in newborns.

1a.7.2. Grade assigned for the quality of the quoted evidence with definition of the grade:

The grades assigned to the quoted evidence for quality are also listed beside the quotes evidence in 1a.4.2.

1a.7.3. Provide all other grades and associated definitions for strength of the evidence in the grading system.
Studies were reviewed and evaluated by ACOG for quality according to the method outlined by the US Preventive Services Task Force:
I Evidence obtained from at least one properly designed randomized controlled trial.
II-1 Evidence obtained from well-designed controlled studies without randomization.
II-2 Evidence obtained from well-designed cohort or case-control analytic studies, preferably from more than one center or research group.
II-3 Evidence obtained from multiple time series with or without the intervention. Dramatic results in uncontrolled experiments also could be regarded as this type of evidence.
III Opinions of respected authorities, based on clinical experience, descriptive studies, or reports of expert committees.

Based on the highest level of evidence found in the data, recommendations are provided and graded by ACOG according to the following categories.

Level A: Recommendations are based on good and consistent Scientific Evidence.
Level B: Recommendations are based on limited or inconsistent Scientific Evidence.
Level C: Recommendations are based primarily on consensus and expert opinion

The CDC used an Evidence-based rating system to determine the strength of their recommendations. Categories, their definitions and strength of recommendations are listed and described in **1a.4.3.**

Ratings and Definitions for the Quality of evidence supporting the recommendations are as follows:
I: Evidence from at least one well-executed randomized, controlled trial or one rigorously designed laboratory-based experimental study that has been replicated by an independent investigator
II: Evidence from at least one well-designed clinical trial without randomization, cohort or case-controlled analytic studies (preferably from more than one center), multiple time-series studies, dramatic results from uncontrolled studies, or some evidence from laboratory experiments
III: Evidence from opinions of respected authorities based on clinical or laboratory experience, descriptive studies, or reports of expert committees.

1a.7.4. What is the time period covered by the body of evidence? (*provide the date range, e.g., 1990-2010*). Date range:
1980-2001: Shoulder Dystocia
1948-2009: Induction of Labor
1967-2009: Management of intrapartum fetal heart tracings
1979-1999: Operative Vaginal Delivery
1966-2010: Prevention of Perinatal Group B Streptococcal Disease

QUANTITY AND QUALITY OF BODY OF EVIDENCE
1a.7.5. How many and what type of study designs are included in the body of evidence? (*e.g., 3 randomized controlled trials and 1 observational study*)
While not required for an outcome measure there is strong evidence for a number of the relationships between clinical practices and our health outcome. For example there are over 20 Randomized Controlled Trials and over 5 large observational studies linking improved neonatal outcomes with GBS screening and prophylactic antibiotics.

Reference: Centers for Disease Control and Prevention. Prevention of Perinatal Group B Streptococcal Disease. MMWR 2010;59 (No. RR 10): 1-36
http://www.cdc.gov/mmwr/pdf/rr/rr5910.pdf

1a.7.6. What is the overall quality of evidence across studies in the body of evidence? (*discuss the certainty or confidence in the estimates of effect particularly in relation to study factors such as design flaws, imprecision due to small numbers, indirectness of studies to the measure focus or target population*)
The evidence varies depending on the clinical practice as to how strongly it impacts the clinical outcome. As noted above, there is very strong evidence for GBS screening and prophylaxis and for proper use of vacuum extractors.

ESTIMATES OF BENEFIT AND CONSISTENCY ACROSS STUDIES IN BODY OF EVIDENCE
1a.7.7. What are the estimates of benefit—magnitude and direction of effect on outcome(s) across studies in the body of evidence? (*e.g., ranges of percentages or odds ratios for improvement/ decline across studies, results of meta-analysis, and statistical significance*)
The magnitude of effect seen in the GBS prevention is quite large->90% reduction in harm. The beneficial effect seen with proper use of vacuum extractors is harder to quantify but several intervention studies have seen 66% reductions in neonatal injuries.

1a.7.8. What harms were studied and how do they affect the net benefit (benefits over harms)?
No harms were noted in these example strategies.

UPDATE TO THE SYSTEMATIC REVIEW(S) OF THE BODY OF EVIDENCE

1a.7.9. If new studies have been conducted since the systematic review of the body of evidence, provide for <u>each</u> new study: 1) citation, 2) description, 3) results, 4) impact on conclusions of systematic review.

Not Applicable

1a.8 OTHER SOURCE OF EVIDENCE

If source of evidence is NOT from a clinical practice guideline, USPSTF, or systematic review, please describe the evidence on which you are basing the performance measure.

n/a

1a.8.1 What process was used to identify the evidence?

n/a

1a.8.2. Provide the citation and summary for each piece of evidence.

n/a

1. Evidence, Performance Gap, Priority – Importance to Measure and Report

Extent to which the specific measure focus is evidence-based, important to making significant gains in healthcare quality, and improving health outcomes for a specific high-priority (high-impact) aspect of healthcare where there is variation in or overall less-than-optimal performance. ***Measures must be judged to meet all subcriteria to pass this criterion and be evaluated against the remaining criteria.***

1a. Evidence to Support the Measure Focus – See attached Evidence Submission Form

Evidence_Submission_Form_UNC_2016-635900134754083149.docx

1b. Performance Gap

Demonstration of quality problems and opportunity for improvement, i.e., data demonstrating:

- considerable variation, or overall less-than-optimal performance, in the quality of care across providers; and/or
- disparities in care across population groups.

1b.1. **Briefly explain the rationale for this measure** *(e.g., the benefits or improvements in quality envisioned by use of this measure)*

The most important childbirth outcome for families is bringing home a healthy baby. While there have been measures developed to assess clinical practices and outcomes in preterm infants, there are a lack of metrics that assess the health outcomes of term infants who represent over 90% of all births.

The Unexpected Complications in Term Newborns metric addresses this gap and measures adverse outcomes resulting in severe or moderate morbidity in otherwise healthy term infants without preexisting conditions. Importantly, this metric also serves as a balancing measure for other NQF endorsed maternal measures such as NTSV Cesarean rates, third and fourth degree lacerations, episiotomies and early elective delivery rates. The purpose of a balancing measure to is guard against any unanticipated or unintended consequences of quality improvement activities for these measures.

1b.2. **Provide performance scores on the measure as specified (current and over time) at the specified level of analysis.** *(This is required for endorsement maintenance. Include mean, std dev, min, max, interquartile range, scores by decile. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included). This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*

2011 data

Dates of data: 1st Jan 2011- 31st Dec 2011

Number of hospitals in California that had births in 2011: n=262

Number of hospitals in California that have eligible cases in the denominator: n=247

Total number of eligible births (included in the denominator): n=392,328

Total number of eligible births in hospitals with over 200 cases in the denominator: n=389,741

Hospital Level Scores (excluding hospitals with less than 200 cases in the denominator N=21):

N=226

Births=389,741

Mean=3.56

Std Dev=2.14

Median=3.01

Min=0.68

Max=15.19

Inter-quartile Range: 25th percentile: 2.16 75th percentile: 4.39

Performance Scores by Deciles (226 hospitals with over 200 cases in denominator)

Performance Score Deciles

Deciles	Number of hospitals	Mean of deciles	StdDev of deciles	Minimum	Maximum
1	22	1.2	0.3	0.7	1.6
2	23	1.8	0.1	1.6	2.0
3	23	2.2	0.1	2.0	2.3
4	22	2.5	0.1	2.4	2.6
5	23	2.9	0.1	2.6	3.0
6	23	3.3	0.2	3.1	3.6
7	22	3.8	0.1	3.6	4.0
8	23	4.3	0.2	4.0	4.7
9	23	5.5	0.4	4.8	6.1

10	22	8.2	2.8	6.2	15.2

2012 data
Date of data: 1st Jan 2012- 31st Dec 2012
Number of hospitals in California that have eligible cases in the denominator in 2012: n=251
Number of hospitals in California that have eligible cases in the denominator in 2012: n=244
Total number of eligible births (included in the denominator): n=397,864
Total number of eligible births in hospitals with over 200 cases in the denominator: n=395,634

Hospital Level Scores (excluding hospitals with less than 200 cases in the denominator N=22)
N=232
Births=395,634
Mean=3.63
Std Dev=2.14
Median=3.14
Min=0.97
Max=14.33
Inter-quartile Range: 25th percentile: 2.18 75th percentile: 4.72
Performance Score Deciles

Deciles	Number of hospitals	Mean of deciles	Std Dev of deciles	Minimum	Maximum
1	23	1.32	0.19	0.97	1.57
2	23	1.75	0.10	1.58	1.95
3	23	2.15	0.10	1.99	2.28
4	24	2.46	0.12	2.29	2.67
5	23	2.93	0.14	2.68	3.13
6	23	3.27	0.09	3.15	3.43
7	24	3.82	0.27	3.45	4.29
8	23	4.65	0.19	4.30	4.89
9	23	5.43	0.42	4.92	6.42
10	23	8.56	2.06	6.42	14.33

1b.3. If no or limited performance data on the measure as specified is reported in 1b2, then provide a summary of data from the literature that indicates opportunity for improvement or overall less than optimal performance on the specific focus of measurement.
Performance Data has been displayed in Question 1.b.2. The data has been produced and tested on over 4 years of administrative data in California. This measure is also in use in Oregon, Washington and by hospitals across the United States who are part of the National Perinatal Information Center´s Network of hospitals.

1b.4. Provide disparities data from the measure as specified (current and over time) by population group, e.g., by race/ethnicity, gender, age, insurance status, socioeconomic status, and/or disability. *(This is required for endorsement maintenance. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities include.) This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*
See tables below for 2011 and 2012 data

2011 data. Includes all 247 eligible hospitals

CHARACTERISTIC	Numerator(Severe + Moderate morbidity)	Denominator	Measure
Total population included in metric	13,666	392,328	3.48%
NEONATAL SEX:			
Male	7,938	200,682	3.96%
Female	5,728	191,646	2.99%
NEONATAL RACIAL/ETHNIC GROUP			
Hispanic, Native Born	3,172	93,691	3.39%

Hispanic, Foreign Born	3,113	96,414	3.23%
Non-Hispanic White	4,119	116,911	3.52%
Non-Hispanic Black	914	21,513	4.25%
Asian / Pacific Islander	1,777	53,040	3.35%
Others	571	10,759	5.31%

NEONATAL INSURANCE STATUS

Medicare-California	6,698	190,885	3.51%
Private	6,041	184,338	3.28%
Self-pay	226	9,303	2.43%
Other	701	7,802	8.98%

NICU LEVEL of HOSPITALS

Basic/No NICU	2307	78,537	2.94%
Intermediate NICU	7492	217,108	3.45%
Community NICU	1842	52,002	3.54%
Regional NICU	2025	44,681	4.53%

GEOGRAPHIC LOCATION of HOSPITALS

Urban	12,306	353,155	3.48%
Non-Urban	1360	39,173	3.47%

HOSPITAL OWNERSHIP

City and/or County & District	2205	56,548	3.90%
For Profit:	1629	55,371	2.94%
Not for Profit/University	9572	273,081	3.51%
Ownership status unknown	260	7328	3.55%

2012 data

CHARACTERISTIC	Numerator	Denominator	Measure
Total	14,178	397,864	3.56%

SEX

Male	8,196	202,804	4.04%
Female	5,981	195,009	3.07%
Other	1	23	4.35%

RACIAL/ETHNIC GROUP

Hispanic, Native Born	3,635	102,865	3.53%
Hispanic, Foreign Born	3,205	95,425	3.36%
Non-Hispanic White	3,952	109,660	3.60%
Non-Hispanic Black	894	20,989	4.26%
Asian / Pacific Islander	1,917	58,473	3.28%
Others	575	10,452	5.50%

INSURANCE STATUS

Medi-Cal	6,864	189,664	3.62%
Private	6,305	186,531	3.38%
Self-pay	276	12,669	2.18%
Other	733	9,000	8.14%

NICU LEVEL of HOSPITAL

Basic NICU	2172	77,231	2.81%
Intermediate NICU	7792	222,014	3.51%
Community NICU	1918	53,141	3.61%
Regional NICU	2296	45,478	5.05%

GEOGRAPHIC LOCATION of HOSPITAL			
Urban	12,666	354,218	3.58%
Non-Urban	1,512	43,646	3.46%
HOSPITAL OWNERSHIP			
City and/or County & District	2387	54,610	4.37%
For Profit:	1530	55,526	2.76%
Not for Profit/University	9998	279,708	3.57%
Ownership status unknown	263	8020	3.28%

1b.5. **If no or limited data on disparities from the measure as specified is reported in 1b4, then provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement. Include citations.**
n/a

1c. High Priority (previously referred to as High Impact)

The measure addresses:

- a specific national health goal/priority identified by DHHS or the National Priorities Partnership convened by NQF; OR
- a demonstrated high-priority (high-impact) aspect of healthcare (e.g., affects large numbers of patients and/or has a substantial impact for a smaller population; leading cause of morbidity/mortality; high resource use (current and/or future); severity of illness; and severity of patient/societal consequences of poor quality).

1c.1. **Demonstrated high priority aspect of healthcare**
Patient/societal consequences of poor quality, Affects large numbers, Frequently performed procedure, High resource use
1c.2. **If Other:**

1c.3. **Provide epidemiologic or resource use data that demonstrates the measure addresses a high priority aspect of healthcare. List citations in 1c.4.**
The birth of a baby is the most common reason for hospitalization in the United States and accounts for over half of all hospital admissions. In 2010, almost 4 million births were reported for US residents (1), with over 3 million meeting our denominator criteria (term newborns without pre-existing conditions). In this population, a high number of procedures are performed to deliver babies, with 32.8% of infants being born via Cesarean section and approximately 3.62% of infants delivered with the assistance of forceps and vacuums (1). This adds up to approximately 1.5 million procedures a year. Barring a prenatal diagnosis, parental expectations for a healthy, normal baby are high with perfection being a common objective.

Since hospitalizations related to childbirth comprise such a large portion of US hospital care, the quality of health care delivered to mothers and their babies should be a focal point for hospitals and health care providers. It is therefore important to quantify and understand the impact of complications for this large population of hospitalized patients (2). Public reporting of hospital outcomes is becoming mainstream (3). Currently, pay for performance is limited to medical/surgical diagnoses, but it is being considered as a mechanism for reimbursement in NICUs and likely will come to bear across all disciplines, including obstetrics (4). In the evolving arena of measuring and reporting quality and patient safety data, it has become the norm to focus on adverse events such as mortality, procedure-related complications, or composite morbidities (3). To assist clinicians and hospitals in improving quality of care we have developed a hospital level performance metric that captures and quantifies neonatal morbidity while accounting for severity of the condition, delivery route (vaginal or cesarean) and neonatal length of stay. We recognize that not all publicly reportable indicators are meaningful or interpretable to the average patient. However, we feel our proposed measure will appeal to both well-informed as well as marginally literate patients who can ask themselves the question “What is the probability that I will go into this hospital and leave with a healthy baby?”

We have significant opportunities to improve care for healthy term infants. Labor, birth management and delivery type can lead to birth injuries, trauma, respiratory complications, hypoxia/asphyxia events and in some infants result in neurologic complications. Neonatal birth injuries can occur due to forceps and vacuum use during delivery. The incidence of subgaleal hematomas following vacuum deliveries is estimated to range from 26 to 45 per 1,000 vacuum deliveries. Other potential neonatal complications associated with vacuum deliveries include intracranial hemorrhage, hyperbilirubinemia, and retinal hemorrhage. The higher rates of neonatal jaundice associated with vacuum delivery may be related to the higher rate of cephalohematoma (5). Brachial plexus injuries and fractures of the clavicle and humerus are associated with shoulder dystocia. The reported incidence of brachial plexus injuries following a delivery complicated by shoulder dystocia varies widely from 4% to 40% most cases of shoulder dystocia resolve without permanent disability making it a moderate morbidity. Fewer than 10% of all cases of shoulder dystocia result in a persistent

brachial plexus injury. Some severe cases of shoulder dystocia may result in hypoxic-ischemic encephalopathy and even death. A study of outcomes from 6,238 cases of shoulder dystocia found that asphyxia was more common among births complicated by shoulder dystocia regardless of maternal diabetic status(6). Sepsis is a leading cause of morbidity in neonates and appropriate screening, diagnositic evaluation and antiobiotic therapy is critical in reducing infection and preventing death (7). Elective delivery between 37 and 39 weeks can lead to respiratory disorders and long NICU stays in some infants, particularly those delivered via Cesarean Section(8-10).

Recent data (including our own, presented in the validity section of this document) have shown that neonatal adverse outcome rates are not the same across the 6-week gestational age range that constitutes term births. Rather, the frequency of adverse outcomes is U-shaped, with the nadir around 39 weeks 0 days though 40 weeks and 6 days gestation (9,11-13). Neonatal morbidities such as respiratory distress syndrome (RDS), ventilator use and neonatal intensive care admissions show the lowest rates between 39 weeks 0 days and 40 weeks 6 days, with higher rates both before 37-38 weeks and after 41-42 weeks(14). Early-term (37 0/7 weeks to 38 6/7 weeks) births are also associated with higher neonatal, postneonatal, and infant mortality rates compared with full-term births (15). The ability to evaluate trends and risks across the spectrum of term gestation is limited for mortality and morbidity. The Unexpected Complications in Term newborns measure addresses this gap by capturing and quantifying severe and moderate neonatal morbidities in term infants that could possibly be reduced or prevented by modifying obstetric and neonatal care practices at the hospital level.

1c.4. Citations for data demonstrating high priority provided in 1a.3

1.Martin JA, Hamilton BE, Ventura SJ et al. Births: Final data for 2010. National vital statistics reports;vol 61 no1. Hyattsville, MD: National Center for health Statistics. 2012

2.Russo, C. A (Thomson Reuters) and Andrews, R.M (AHRQ). Potentially Avoidable Injuries to mothers and Newborns During Childbirth, 2006. HCUP Statistical Brief # 74. June 2009. Agency for Healthcare Research and Quality. Rockville, MD. http://www.hcup-us.ahrq.gov/reports/statsbriefs/sb74.pdf.

3.Gregory KD, Fridman M, Shah S et al. Global measures of quality and patient safety-related childbirth outcomes: should we monitor adverse or ideal rates? Am J Obstet Gynecol 2009;200:681.e1-681.e7.

4.Profit J, Zupancic JA, Gould JB et al. Implementing pay-for-performance in the neonatal intensive care unit. Pediatrics 2007;119:975-82

5.Operative Vaginal Delivery. Practice Bulletin No. 17. American College of Obstetricians and Gynecologists. Obstet Gynecol 2000;1-8

6.Shoulder dystocia. ACOG Practice Bulletin No. 40. American College of Obstetrician and Gynecologists. Obstet Gynecol 2002;100:1045-50

7. Centers for Disease Control and Prevention. Prevention of Perinatal Group B Streptococcal Disease. MMWR 2010;59 (No. RR 10): 1-36 http://www.cdc.gov/mmwr/pdf/rr/rr5910.pdf

8.Wilmink FA, Hukkelhoven CW, Lunshof, S et al. Neonatal outcome following elective cesarean section beyond 37 weeks of gestation: a 7- year retrospective analysis of a national registry.2010 Am J Obstet Gynecol 2002:250.e1-8

9. Tita AT, Landon MB, Spong CY et al. Timing of Elective Repeat Cesarean Delivery at Term and Neonatal Outcomes. N Engl J Med 2009;360(2):111-20

10.Hansen AK, Wisborg K, Uldbjerg N. Risk of respiratory morbidity in term infants delivered by elective cesarean section: cohort study. BMJ 2008;336:85

11.Spong CY. Defining "Term" pregnancy: Recommendations from the defining "term" pregnancy workgroup. JAMA. 2013 Jun 19;309(23):2445-6.

12. Zhang X and Kramer MS. Variations in Mortality and Morbidity by Gestational Age among Infants Born at Term. J Pediatr 2009;154:358-62

13. Fleischman AR, Oinuma M and Clark SL. Rethinking the Definition of "Term Pregnancy". Obstet Gynecol 2010;116(1)136-139

14.Clark SL, Miller DD, Belfort MA, et al. Neonatal and maternal outcomes associated with elective term delivery. Am J Obstet Gynecol 2009;200:156.e1-156.e4

15.Reddy UM, Bettegowda VR, Dias T et al. Term Pregnancy: A period of Heterogeneous Risk for Infant Mortality. Obstet Gynecol 2011;117:1279-1287

1c.5. If a PRO-PM (e.g. HRQoL/functional status, symptom/burden, experience with care, health-related behaviors), provide evidence that the target population values the measured PRO and finds it meaningful. *(Describe how and from whom their input was obtained.)*

Not Applicable

2. Reliability and Validity—Scientific Acceptability of Measure Properties

Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. ***Measures must be judged to meet the subcriteria for both reliability and validity to pass this criterion and be evaluated against the remaining criteria.***

2a.1. Specifications The measure is well defined and precisely specified so it can be implemented consistently within and across organizations and allows for comparability. eMeasures should be specified in the Health Quality Measures Format (HQMF) and the Quality Data Model (QDM).

De.5. **Subject/Topic Area** *(check all the areas that apply):*
Perinatal and Reproductive Health, Perinatal and Reproductive Health : Newborn

De.6. **Cross Cutting Areas** *(check all the areas that apply):*
Safety : Complications

S.1. **Measure-specific Web Page** *(Provide a URL link to a web page specific for this measure that contains current detailed specifications including code lists, risk model details, and supplemental materials. Do not enter a URL linking to a home page or to general information.)*
https://www.cmqcc.org/focus-areas/quality-metrics/unexpected-complications-term-newborns

S.2a. **If this is an eMeasure**, HQMF specifications must be attached. Attach the zipped output from the eMeasure authoring tool (MAT) - if the MAT was not used, contact staff. (Use the specification fields in this online form for the plain-language description of the specifications)
This is not an eMeasure **Attachment:**

S.2b. **Data Dictionary, Code Table, or Value Sets** *(and risk model codes and coefficients when applicable) must be attached. (Excel or csv file in the suggested format preferred - if not, contact staff)*
Attachment **Attachment:** Unexpected_Newborn_Complications_Appendices-635908840574237076.xlsx

S.3. **For endorsement maintenance**, please briefly describe any changes to the measure specifications since last endorsement date and explain the reasons.
Since the initial NQF endorsement there has been useful discussion and feedback by Obstetric and Neonatal leaders around the nation about how to make this measure more useful for quality improvement. Likewise we have had feedback from patient user groups as to how to make it more understandable from their perspective. A summary of our subsequent changes and updates is presented below.
•Change in Name: The initial name, Healthy Term Newborn, was chosen to represent a positive health status but many women had a hard time understanding what we were measuring. We chose to invert the measure to Unexpected Newborn Complications for that reason and because it was easier to interpret the typical range from 1 to 6% than from 94 to 99% (all of which sound good).
•Adding two sub--measures (Severe and Moderate): A number of experts were concerned that the scope of the measure was too broad, covering conditions of varying severity that may not normally go together. After several user groups and expert panels we broke the ICD-9 codes into the two categories to allow analysis by severity but to also allow roll-up for summary assessment.
•Identification of Term Infants: As the measure was put into production, it was surprisingly hard to identify term infants using Patient Discharge Diagnosis (PDD) data alone as called for in the initial specifications. Available ICD-9 codes for gestational age and birth weight are significantly under-utilized in practice. However, the use of a linked PDD to Birth Certificate data files has solved this issue as the Birth Certificate data fields for Best Obstetric Gestational Age and Birth weight have high degrees of completeness and accuracy. The Birth Certificate also allows the addition of extremely low Apgar scores as an additional numerator item.
•Additional Codes for Denominator Exclusions: After feedback was received on additional potential congenital anomaly ICD-9 codes, we added a wider panel of codes so that now essentially any congenital malformation is excluded. (see Appendix 2)
•Coding Practices: Checks and balances: We learned that coding practices do vary for some of these codes with some hospitals being “over exuberant” in their coding and others clearly under-coding existing complications. The new specifications attempt to balance this issue by requiring that many codes for Moderate Complications additionally have an infant length of stay that exceeds the typical maternal postpartum length of stay (greater than 2 days for a vaginal birth and greater than 4 days for a cesarean birth). This requirement significantly reduces the number of infants identified but validates that these babies had significant morbidity. Conversely, some babies had very long neonatal length of stay without any codes to account for it, suggesting the possibility of under-coding. Our expert panel identified two categories of prolonged neonatal length of stay that were not medically serious and could be excluded from this consideration: neonatal jaundice typically treated with Billi-Lights, and social disruption for homelessness or foster care. (see Appendices 4 and 5)

•Septicemia Code: We found that a number of babies with septicemia (771.81) had short length of stay indicating that it was not likely to be real sepsis (in fact more likely to be "r/o sepsis"), therefore we added a requirement for a length of stay of at least 5 days to be included among Severe Complications

S.4. Numerator Statement *(Brief, narrative description of the measure focus or what is being measured about the target population, i.e., cases from the target population with the target process, condition, event, or outcome)*
IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
Numerator: The numerator is divided into two categories: Severe complications and moderate complications.
Severe complications include neonatal death, transfer to another hospital for higher level of care, extremely low Apgar Scores (=3 at either 5 or 10 minutes of life), severe birth injuries such as intracranial hemorrhage or nerve injury, neurologic damage, severe respiratory and infectious complications such as sepsis. Parents of such babies may often worry about short or long term infant outcomes.
Moderate complications include diagnoses or procedures that raise concern but at a lower level than the list for severe (e.g. use of CPAP or bone fracture). For inclusion in the numerator, most require an infant length of stay that exceeds that of the mother, validating that these are indeed significant complications. Examples include less severe respiratory complications (e.g. Transient Tachypnea of the Newborn), or infections with a longer length of stay not including sepsis. As a "safety net" to capture cases who were under-coded, the numerator also includes infants who have a prolonged length of stay of over 5 days to capture the "seemingly normal" infants with neither any form of jaundice nor a social reason for staying in the hospital (e.g. family disruption or adoption).

S.5. Time Period for Data *(What is the time period in which data will be aggregated for the measure, e.g., 12 mo, 3 years, look back to August for flu vaccination? Note if there are different time periods for the numerator and denominator.)*
The data was looked at for the birth hospitalization only. The data was analyzed for the 2011 and 2012 calendar years (separately). Results can also be tracked monthly and quarterly.

S.6. Numerator Details *(All information required to identify and calculate the cases from the target population with the target process, condition, event, or outcome such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
IF an OUTCOME MEASURE, describe how the observed outcome is identified/counted. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
In the full term neonatal population that excluded premature infants, low birth weight babies, infants with congenital malformations, fetuses with pre-existing conditions such as IUGR and babies exposed to maternal drug use, babies were selected for inclusion in the numerator in a hierarchical manner as follows:

PART A: Severe Complications: Identify and include the following in a hierarchical manner:
a) Neonatal Deaths (Use patient discharge diagnosis data, specifically the disposition code for death)
b) Neonatal Transfers (Use patient discharge diagnosis data, specifically the disposition code for transfer to a higher level of care)
c) Low Apgar Scores at 5 minutes or 10 minutes of <=3 out of a possible 10 (Use Birth certificate to obtain Apgar scores)
d) Severe Morbidities: (Use patient discharge diagnosis data, examining both primary and other diagnosis and procedure fields for ICD-9 Codes defining an array of specific severe complications. Please refer to Appendix 3, Groups 3A through 3I as the codes are too numerous to include here)
e) Sepsis with a neonatal Length of Stay that exceeds 4 days (Use patient discharge diagnosis data, examining both primary and other diagnosis fields for the specific ICD-9 code defining sepsis. Note that neonatal stay is defined as the date of discharge minus the date of birth).

The neonates identified in Part A make up the "Severe Complications" component of the numerator.
In the remaining infants (those without severe morbidities), identify and include the following

PART B: Moderate Complications: Identify and include the following in a hierarchical manner:
a) Moderate complications not requiring a specific length of stay: Identify babies with moderate complications that do not require a specific length of stay for inclusion (Use Patient discharge Diagnosis data, examining both primary and other diagnosis and procedure fields for ICD-9 codes identifying specific moderate complications (see Appendix 4, Groups A though C as the codes are too numerous to include here)
b) Specific Prolonged neonatal length of Stay stratified by method of delivery. Among babies who were delivered vaginally, identify those who have a length of stay of over 2 days. Among babies delivered via Cesarean Section, identify those who have a length of stay of over 4 days. (Use V-code 30.00 to identify vaginal births, and V30.01 to identify Cesarean births. V-codes are found in patient

discharge data. Neonatal length of stay is defined as the date of discharge minus the date of birth).
c) Moderate complications requiring a prolonged length of stay: Among the infants identified in step b, identify those with moderate complications (Use Patient discharge Diagnosis data, examining both primary and other diagnosis and procedure fields for ICD-9 codes identifying specific moderate complications that require a prolonged length of stay for inclusion in the numerator. See Appendix 4, Groups D through H)
d) Prolonged neonatal Length of Stay that Exceeds 5 days: In the remaining population, identify babies who have a prolonged length of stay that exceeds 5 days. (Use Patient Discharge Diagnosis Data to determine Length of Stay. Neonatal length of stay is defined as the date of discharge minus the date of birth).
e) Exclude infants with jaundice or social indications: Among babies identified as having a length of stay that exceeds 5 days, exclude those who have jaundice or are in hospital for social indications such as adoption or foster care. (See Appendix 5 on our web-page for jaundice and social exclusion codes)

S.7. **Denominator Statement** *(Brief, narrative description of the target population being measured)*
The denominator is comprised of singleton, live born babies who are at least 37.0 weeks of gestation, and over 2500g in birth weight. The denominator excludes most serious fetal conditions that are "preexisting" (present before labor), including prematurity, multiple gestations, poor fetal growth, congenital malformations, genetic disorders, other specified fetal and maternal conditions and infants exposed to maternal drug use in-utero. The final denominator population consists of babies who are expected to do well following labor and delivery and go home routinely with their mothers.

S.8. **Target Population Category** *(Check all the populations for which the measure is specified and tested if any):*
Children's Health, Maternal Health

S.9. **Denominator Details** *(All information required to identify and calculate the target population/denominator such as definitions, specific data collection items/responses , code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
Step 1: Identify and include singleton, inborn, live births (Use Patient discharge Diagnosis data, specifically diagnosis Codes V30.00 or V30.01).
Step 2: Identify and include babies with birth weight >= 2500g. (Use birth certificate or Patient Discharge data).
Step 3: Identify and include full term babies, >=37 weeks gestation (Use birth certificate variable called best obstetric estimate of gestational age).
Step 4: In less than 1% of cases, the best obstetric estimate of gestation age is missing. In these cases, use LMP-based gestational age to identify full term infants. (Use birth certificate or Patient Discharge data).
Step 5: If both sources of gestational age are missing, include only infants who are over 3000g, as they are more likely to be full term.
Step 6: In the singleton, full term, population obtained in steps 1 through 5, identify and exclude babies with all congenital malformations and genetic disorders (See Appendix 2, Group A for the list of congenital malformation and genetic disorder exclusions)
Step 7: After congenital malformations and genetic disorders are excluded, further exclude babies with fetal conditions such as IUGR (see Appendix 2, Group B for the list of preexisting fetal conditions to be excluded)
Step 8: After babies with congenital malformations, genetic disorders and fetal conditions are excluded, further exclude infants who were exposed to maternal drug use in-utero. (see Appendix 2, Group C for the list of maternal drug use exposures to be excluded)
**Note: List of ICD-9 codes with individual descriptors is available in the Appendices on our web-page

S.10. **Denominator Exclusions** *(Brief narrative description of exclusions from the target population)*
a) Babies not born in hospitals are excluded as this is a hospital quality performance measure
b) Babies who are part of multiple gestation pregnancies are excluded.
c) Premature infants (babies born before 37 weeks gestational age) are excluded
d) Low birth weight babies (<=2500g) are excluded
e) Babies with congenital malformations and genetic diseases are excluded
f) Babies with pre-existing fetal conditions such as IUGR are excluded
g) Babies who were exposed to maternal drug use in-utero are excluded

S.11. **Denominator Exclusion Details** *(All information required to identify and calculate exclusions from the denominator such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
a)Babies not born in hospitals are excluded as this is a hospital quality performance measure (Exclude all other live birth codes other

than V30.00 and V30.01)
b)Babies who are part of multiple gestation pregnancies are excluded. (Exclude all other live birth codes other than V30.00 and V30.01)
c)Premature infants (babies born before 37 weeks gestational age) are excluded (use best obstetric estimate of gestational age found in the birth certificate to exclude all infants born before 37 weeks. If best obstetric of gestational age is missing, use the LMP gestational age variable instead to identify infants under 37 weeks)
d)Low birth weight babies (<=2500g) are excluded (Use birth certificate birth weight variable to identify infants under 2500g)
e)Babies with congenital malformations and genetic diseases are excluded (Use ICD-9 codes listed in Appendix 2, Group A to exclude infants with these conditions)
f)Babies with pre-existing fetal conditions such as IUGR are excluded (Use ICD-9 codes listed in Appendix 2, Group B to exclude infants with these conditions)
g)Babies who were exposed to maternal drug use in-utero are excluded (Use ICD-9 codes listed in Appendix 2, Group C to exclude infants with these conditions)

S.12. Stratification Details/Variables *(All information required to stratify the measure results including the stratification variables, definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format with at S.2b)*
Not applicable

S.13. Risk Adjustment Type (Select type. Provide specifications for risk stratification in S.12 and for statistical model in S.14-15)
No risk adjustment or risk stratification
If other:

S.14. Identify the statistical risk model method and variables *(Name the statistical method - e.g., logistic regression and list all the risk factor variables. Note - risk model development and testing should be addressed with measure testing under Scientific Acceptability)*
None

S.15. Detailed risk model specifications *(must be in attached data dictionary/code list Excel or csv file. Also indicate if available at measure-specific URL identified in S.1.)*
Note: Risk model details (including coefficients, equations, codes with descriptors, definitions), should be provided on a separate worksheet in the suggested format in the Excel or csv file with data dictionary/code lists at S.2b.

S.15a. Detailed risk model specifications (*if not provided in excel or csv file at S.2b*)
Not applicable

S.16. Type of score:
Rate/proportion
If other:

S.17. Interpretation of Score *(Classifies interpretation of score according to whether better quality is associated with a higher score, a lower score, a score falling within a defined interval, or a passing score)*
Better quality = Lower score

S.18. Calculation Algorithm/Measure Logic *(Describe the calculation of the measure score as an ordered sequence of steps including identifying the target population; exclusions; cases meeting the target process, condition, event, or outcome; aggregating data; risk adjustment; etc.)*
STEP 1: Calculate Denominator Inclusions

a)Identify and include singleton, inborn, live births (Use Patient discharge Diagnosis data, specifically diagnosis Codes V30.00 or V30.01 listed in Appendix 1).
b)Next, identify and include babies with birth weight >= 2500g. (Use birth certificate or Patient Discharge data).
c)Next, identify and include full term babies, >=37 weeks gestation (Use birth certificate variable called best obstetric estimate of gestational age). In less than 1% of cases, the best obstetric estimate of gestation age is missing. In these cases, use LMP-based gestational age to identify full term infants. (Use birth certificate or Patient Discharge data).
d)If both sources of gestational age are missing, include only infants who are over 3000g, as they are more likely to be full term. (Use the birth certificate variable for birth weight).

STEP 2: Calculate Denominator Exclusions

a)In the singleton, full term, population of neonates obtained in Step 1, identify and exclude babies with all congenital malformations and genetic disorders (Use codes listed in Appendix 2, Group A to exclude infants)
b)After congenital malformations and genetic disorders are excluded, further exclude babies with fetal conditions such as IUGR (Use codes listed in Appendix 2, Group B to exclude infants)
c)After babies with congenital malformations, genetic disorders and fetal conditions are excluded, further exclude infants who were exposed to maternal drug use in-utero. (Use codes listed in Appendix 2, Group C to exclude infants).
d)This is the measure's final denominator population

Step 3: Numerator Inclusions: PART A: SEVERE COMPLICATIONS

a)Identify and include Neonatal Deaths (Using patient discharge diagnosis data, specifically the disposition code for death)
b)Identify and include neonatal transfers (Using patient discharge diagnosis data, specifically the disposition code for transfer to a higher level of care)
c)Identify and include babies with "Apgar at 5 minutes" OR "Apgar at 10 minutes" scores of less than 4 (Use Birth certificate or medical record to obtain Apgar scores)
d)Identify and include babies with Severe Morbidities (Use patient discharge diagnosis data, examining both primary and other diagnosis and procedure fields for specific ICD-9 Codes defining an array of specific severe complications. Please refer to Appendix 3, Groups 3A through 3I as the codes are too numerous to include here)
e)Identify and include babies with a Sepsis code and a length of stay that exceeds 4 days (Use patient discharge diagnosis data, examining both primary and other diagnosis fields for the specific ICD-9 code defining sepsis but also requiring a neonatal length of stay of over 4 days. Note that neonatal stay is defined as the date of discharge minus the date of birth).

The neonates identified in Step 3 comprise the "Severe Complications" component of the numerator.

Step 4: Numerator Inclusions: PART B: MODERATE COMPLICATIONS
In the remaining infants (those without severe morbidities), identify and include the following

a)Identify babies with moderate complications that do not require a specific length of stay for inclusion (Use Patient discharge Diagnosis data, examining both primary and other diagnosis and procedure fields for specific ICD-9 codes identifying specific moderate complications (see Appendix 4, Groups A though C)
b)Identify babies with a specified prolonged length of stay stratified by method of delivery. In the population of babies who were delivered vaginally, identify those who have a length of stay of over 2 days. Among babies delivered via Cesarean Section, identify those who have a length of stay of over 4 days.
c)Among babies identified as having a prolonged length of stay (stratified by method of delivery), identify and include those who have moderate complications (Use Patient discharge Diagnosis data, examining both primary and other diagnosis and procedure fields for specific ICD-9 codes identifying specific moderate complications. See Appendix 4, Groups D through H)
d)In the remaining population, identify babies who have a prolonged length of stay that exceeds 5 days. Use Patient Discharge Diagnosis Data to determine Length of Stay
e)Among babies identified as having a length of stay that exceeds 5 days, exclude those who have jaundice or are in hospital for social indications such as adoption or foster care (See Appendix 5 for jaundice and social exclusion codes)

Step 5: Calculation of Unexpected Complications in Term Newborns measure:

Unexpected Newborn Complications (Total): Rate per 100 live births.
(Severe Complications + Moderate Complications/ Final Denominator) x100

S.19. **Calculation Algorithm/Measure Logic Diagram URL or Attachment** *(You also may provide a diagram of the Calculation Algorithm/Measure Logic described above at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
Available in attached appendix at A.1

S.20. **Sampling** *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*
IF a PRO-PM, identify whether (and how) proxy responses are allowed.

Not applicable. The measure does not require sampling or a survey. However, it is recommended that hospitals have at least 200 qualifying cases in the denominator population of this metric (i.e) Full term infants with no pre-existing conditions, malformations, etc described in S.9

S.21. **Survey/Patient-reported data** *(If measure is based on a survey, provide instructions for conducting the survey and guidance on minimum response rate.)*
IF a PRO-PM, specify calculation of response rates to be reported with performance measure results.
Not applicable

S.22. **Missing data** (specify how missing data are handled, e.g., imputation, delete case.)
Required for Composites and PRO-PMs.
This measure has several methods to deal with missing data, under coding and over coding:

HOW THIS MEASURE HANDLES MISSING DATA IN KEY VARIABLES:
1)Missing V Codes: V codes 30.00 and 30.01 identify live, inborn vaginal and cesarean births. If a case does not have one of these two codes, it is excluded from the calculation of the metric.

2)Missing Gestational Age: Best Obstetric Estimate of gestational age is a key variable in the calculation of this measure. It is available on the birth certificate in over 99% of the records. To account for missing gestational age in less than 1% of records, there are 2 extra steps in the algorithm that allow hospitals to use an alternate gestational age instead, (gestational age based on last menstrual period). If BOTH sources of gestational age are missing, a birth weight of 3000g and above is required for inclusion of the case in the measure calculation.

3)Missing Birth Weight: If birth weight is missing from the birth certificate, a case can be included if the baby is 37 or over weeks of gestation. If both birth weight and gestational age are missing, a case is excluded from the measure calculation. This occurred in only 107 cases out of over 500,000 births in our study groups.

4)Missing Apgar Scores (5 or 10 minutes): If a newborn had a score of less than 4 on his/her Apgar tests (at 5 or 10 minutes) and this is not recorded on the birth certificate, this infant would be captured in later steps of the algorithm where the severe or moderate morbidity diagnosis and procedure codes or prolonged length of stay would ensure that he/she was included in the numerator.

5)Deaths/Transfers: These are initially captured using the discharge codes but will also be captured using the ICD9 codes.

HOW THIS MEASURE HANDLES UNDERCODING OR OVERCODING:
6)Under coding of diagnoses or procedure codes: If a baby is missing a diagnosis code (e.g HIE/asphyxia), the case is still likely to be identified using procedure codes (head cooling, intubation, resuscitation) or a neonatal length of stay that exceeds the standard length of stay that a mother is allowed postpartum (2 days for vaginal births and 4 days for cesarean sections).

7)Over coding of certain diagnoses or procedure codes: Certain codes may not truly represent serious morbidity such as Sepsis with a neonatal length of stay that is under 4 days. This may most likely be “rule out sepsis” rather than true sepsis. Another example is the general “catch all” code of “other birth injuries” that do not keep the baby in hospital longer than the mother.

8)Prolonged Neonatal Length of Stay: The final step in the measure algorithm is a neonatal length of stay that exceeds 5 days and is not due to neonatal jaundice or social indicators such as adoption or foster care. This step captures infants whose diagnoses or procedures may be under coded or missing.

S.23. **Data Source** *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*
If other, please describe in S.24.
Administrative claims

S.24. **Data Source or Collection Instrument** *(Identify the specific data source/data collection instrument e.g. name of database, clinical registry, collection instrument, etc.)*
IF a PRO-PM, identify the specific PROM(s); and standard methods, modes, and languages of administration.
This measure utilizes a linked dataset obtained from two separate data sources, patient discharge data and birth certificate files.
Patient Discharge Data:
Obtained from the Office of Statewide Planning and Discharge (OSHPD). This dataset does not include data on births from

military/naval hospitals as they do not submit data to OSHPD.

Linked to:
Birth Certificate Files:
Obtained from the Center for Health Statistics

S.25. **Data Source or Collection Instrument** *(available at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*

S.26. **Level of Analysis** *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*
Facility, Integrated Delivery System, Population : Regional, Population : State

S.27. **Care Setting** *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*
Hospital/Acute Care Facility
If other:

S.28. **COMPOSITE Performance Measure** - Additional Specifications *(Use this section as needed for aggregation and weighting rules, or calculation of individual performance measures if not individually endorsed.)*
Not applicable. This is not a composite measure.

2a. Reliability – See attached Measure Testing Submission Form
2b. Validity – See attached Measure Testing Submission Form
Measure_testing_form_template_Feb2016-635920863038793157.docx

NATIONAL QUALITY FORUM—Measure Testing (subcriteria 2a2, 2b2-2b7)

Measure Number (*if previously endorsed*)**:** Click here to enter NQF number

Measure Title: Click here to enter measure title

Date of Submission: Click here to enter a date

Type of Measure:

☐ Composite – ***STOP – use composite testing form***	☒ Outcome (*including PRO-PM*)
☐ Cost/resource	☐ Process
☐ Efficiency	☐ Structure

Instructions

- Measures must be tested for all the data sources and levels of analyses that are specified. ***If there is more than one set of data specifications or more than one level of analysis, contact NQF staff*** about how to present all the testing information in one form.
- **For all measures, sections 1, 2a2, 2b2, 2b3, and 2b5 must be completed.**
- **For outcome and resource use measures**, section **2b4** also must be completed.
- If specified for **multiple data sources/sets of specificaitons** (e.g., claims and EHRs), section **2b6** also must be completed.
- Respond to all questions as instructed with answers immediately following the question. All information on testing to demonstrate meeting the subcriteria for reliability (2a2) and validity (2b2-2b6) must be in this form. An appendix for *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 20 pages (*incuding questions/instructions;* minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.
- For information on the most updated guidance on how to address sociodemographic variables and testing in this form refer to the release notes for version 6.6 of the Measure Testing Attachment.

Note: The information provided in this form is intended to aid the Steering Committee and other stakeholders in understanding to what degree the testing results for this measure meet NQF's evaluation criteria for testing.

2a2. Reliability testing [10] demonstrates the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise. For **PRO-PMs and composite performance measures**, reliability should be demonstrated for the computed performance score.

2b2. Validity testing [11] demonstrates that the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality. For **PRO-PMs and composite performance measures**, validity should be demonstrated for the computed performance score.

2b3. Exclusions are supported by the clinical evidence; otherwise, they are supported by evidence of sufficient frequency of occurrence so that results are distorted without the exclusion; [12]
AND
If patient preference (e.g., informed decisionmaking) is a basis for exclusion, there must be evidence that the exclusion impacts performance on the measure; in such cases, the measure must be specified so that the information about patient preference and the effect on the measure is transparent (e.g., numerator category computed separately, denominator exclusion category computed separately). [13]

2b4. For outcome measures and other measures when indicated (e.g., resource use):

- **an evidence-based risk-adjustment strategy** (e.g., risk models, risk stratification) is specified; is based on patient factors (including clinical and sociodemographic factors) that influence the measured outcome and are present at start of care; [14,15] and has demonstrated adequate discrimination and calibration

OR

- rationale/data support no risk adjustment/ stratification.

2b5. Data analysis of computed measure scores demonstrates that methods for scoring and analysis of the specified measure allow for **identification of statistically significant and practically/clinically meaningful [16] differences in performance**;

OR

there is evidence of overall less-than-optimal performance.

2b6. If multiple data sources/methods are specified, there is demonstration they produce comparable results.

2b7. For **eMeasures, composites, and PRO-PMs** (or other measures susceptible to missing data), analyses identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias.

Notes

10. Reliability testing applies to both the data elements and computed measure score. Examples of reliability testing for data elements include, but are not limited to: inter-rater/abstractor or intra-rater/abstractor studies; internal consistency for multi-item scales; test-retest for survey items. Reliability testing of the measure score addresses precision of measurement (e.g., signal-to-noise).

11. Validity testing applies to both the data elements and computed measure score. Validity testing of data elements typically analyzes agreement with another authoritative source of the same information. Examples of validity testing of the measure score include, but are not limited to: testing hypotheses that the measures scores indicate quality of care, e.g., measure scores are different for groups known to have differences in quality assessed by another valid quality measure or method; correlation of measure scores with another valid indicator of quality for the specific topic; or relationship to conceptually related measures (e.g., scores on process measures to scores on outcome measures). Face validity of the measure score as a quality indicator may be adequate if accomplished through a systematic and transparent process, by identified experts, and explicitly addresses whether performance scores resulting from the measure as specified can be used to distinguish good from poor quality.

12. Examples of evidence that an exclusion distorts measure results include, but are not limited to: frequency of occurrence, variability of exclusions across providers, and sensitivity analyses with and without the exclusion.

13. Patient preference is not a clinical exception to eligibility and can be influenced by provider interventions.

14. Risk factors that influence outcomes should not be specified as exclusions

15. With large enough sample sizes, small differences that are statistically significant may or may not be practically or clinically meaningful. The substantive question may be, for example, whether a statistically significant difference of one percentage point in the percentage of patients who received smoking cessation counseling (e.g., 74 percent v. 75 percent) is clinically meaningful; or whether a statistically significant difference of $25 in cost for an episode of care (e.g., $5,000 v. $5,025) is practically meaningful. Measures with overall less-than-optimal performance may not demonstrate much variability across providers.

1. DATA/SAMPLE USED FOR ALL TESTING OF THIS MEASURE

Often the same data are used for all aspects of measure testing. In an effort to eliminate duplication, the first five questions apply to all measure testing. If there are differences by aspect of testing,(e.g., reliability vs. validity) be sure to indicate the specific differences in question 1.7.

1.1. What type of data was used for testing? (*Check all the sources of data identified in the measure specifications and data used for testing the measure*. *Testing must be provided for all the sources of data specified and intended for measure implementation.* ***If different data sources are used for the numerator and denominator, indicate N [numerator] or D [denominator] after the checkbox.***)

Measure Specified to Use Data From: (*must be consistent with data sources entered in S.23*)	Measure Tested with Data From:
☐ abstracted from paper record	☐ abstracted from paper record
☐ administrative claims	☒ administrative claims
☐ clinical database/registry	☐ clinical database/registry
☐ abstracted from electronic health record	☐ abstracted from electronic health record
☐ eMeasure (HQMF) implemented in EHRs	☐ eMeasure (HQMF) implemented in EHRs
☐ other: Click here to describe	☐ other: Click here to describe

1.2. If an existing dataset was used, identify the specific dataset (*the dataset used for testing must be consistent with the measure specifications for target population and healthcare entities being measured; e.g., Medicare Part A claims, Medicaid claims, other commercial insurance, nursing home MDS, home health OASIS, clinical registry*).

2011: This measure has been tested on California discharge data sets for several years (2004 to 2007) with ~560,000 births per year. We also examined these codes on the HCUP data set that comprised over 8 million births.

The linked dataset identified in S.24.is the same that has been used in all aspects of this measure including testing.

Patient Discharge Data: Obtained from the Office of Statewide Planning and Discharge (OSHPD), State of California. OSHPD datasets do not include data on births from military hospitals. Linked to **Birth Certificate Files:** Obtained from California Department of Public Health: Center for Health Information and Statistics.

The additional dataset used for Validity testing were the National Perinatal Center's (NPIC) Perinatal Center Data Base.

1.3. What are the dates of the data used in testing? Click here to enter date range

Dates of testing were January 1, 2011- December 31, 2011

1.4. What levels of analysis were tested? (*testing must be provided for <u>all</u> the levels specified and intended for measure implementation, e.g., individual clinician, hospital, health plan*)

Measure Specified to Measure Performance of: (*must be consistent with levels entered in item S.26*)	Measure Tested at Level of:
☐ individual clinician	☐ individual clinician
☐ group/practice	☐ group/practice
☒ hospital/facility/agency	☒ hospital/facility/agency
☐ health plan	☐ health plan
☒ other: regional, integrated delivery system, state	☒ other: regional, integrated delivery system, state

1.5. How many and which <u>measured entities</u> were included in the testing and analysis (by level of analysis and data source)? (*identify the number and descriptive characteristics of measured entities included in the analysis (e.g., size, location, type); if a sample was used, describe how entities were selected for inclusion in the sample*)

There were 262 hospitals in California with recorded births in 2011. Of these 262 hospitals, 15 did not have any eligible cases that met our denominator inclusion criteria (full term infants with no preexisting conditions) and were thus excluded from further analysis.
Listed below are the characteristics of the 247 hospitals that met our inclusion criteria and were therefore included in testing and analysis unless specified otherwise.

2011 data

Geographic Location (Urban/Rural)
Urban: 195
Rural: 52

Neonatal Intensive Care Unit Type
Basic NICU: 101
Community NICU: 27
Intermediate NICU: 101
Regional NICU: 18

Volume of Births:
Under 1000 births annually: 79
1000-3000 births annually: 117
Over 3000 births annually: 51

Hospital Ownership Type
City and/or County/District: 45
For Profit hospitals: 42
Non-profit corporation/University: 151
Unknown: 9

1.6. How many and which patients were included in the testing and analysis (by level of analysis and data source)? (*identify the number and descriptive characteristics of patients included in the analysis (e.g., age, sex, race, diagnosis); if a sample was used, describe how patients were selected for inclusion in the sample*)

Testing was performed on 2011 data. A total of 392,328 babies from 247 hospitals were included in the analysis. Singleton term newborns without preexisting conditions included in the testing and analysis are as follows:

2011 data

	Number of babies	**% of final denominator**
Gestational Age		
≤ 37 weeks GA:	27,989	99.94
GA unknown but identified by appropriate birth weight of over 3000g	224	0.06
Sex:		
Males: 200,682	200,682	51.15
Females: 191,646	191,646	48.85
Race:		
Hispanic, Native Born	93,691	23.9
Hispanic, Foreign Born	96,414	24.6
Non-Hispanic White	116,911	29.8
Non-Hispanic Black	21,513	5.5
Asian/Pacific Islander	53,040	13.5
Other	10,759	2.74

Method of Delivery		
Cesarean Section	120,967	30.83
Vaginal birth	271,361	69.17

1.7. If there are differences in the data or sample used for different aspects of testing (e.g., reliability, validity, exclusions, risk adjustment), identify how the data or sample are different for each aspect of testing reported below.

Reliability Sample: Reliability testing was performed on 2011 data. Excluding hospitals (n=21) that contributed fewer than 200 cases to the denominator, the final testing sample consisted of 226 hospitals and 387,677 babies who met the denominator's inclusion criteria.

Validity Sample 1: For the Empirical validity portion, testing was performed using National Perinatal Information Center (NPIC)'s Perinatal Center Data Base. The sample was limited to babies with gestational ages 37 weeks and 0 days to 42 weeks and 6 days and consisted of 174,501 cases for the year 2011. The gestational age variable used was Best Obstetric Estimate of Gestational Age. The National Perinatal Information Center (NPIC)'s Perinatal Center Data Base is comprised primarily of hospitals offering subspecialty perinatal services and includes academic and non-academic institutions. NPIC's 73 member hospitals from across the country contribute approximately 725,000 discharges annually making it one of the largest repositories for hospital based perinatal clinical and financial discharge data in the United States. The large sample size contains a combination of clinical and financial information that enables users to perform linked analyses on maternal complications/comorbidities and infant outcomes.

The **second set of empirical validity testing** for the Unexpected Newborn Complications metric **serving as a balancing measure for NTSV cesarean section,** was performed for a separate initiative undertaken by the California Maternal Quality Care Collaborative (CMQCC) and pilot hospitals using data from the California Maternal Data Center.

1.8 What were the patient-level sociodemographic (SDS) variables that were available and analyzed in the data or sample used? For example, patient-reported data (e.g., income, education, language), proxy variables when SDS data are not collected from each patient (e.g. census tract), or patient community characteristics (e.g. percent vacant housing, crime rate).

Patient level sociodemographic (SDS) variables that are available in the CMQCC linked dataset we used are maternal age, race/ethnicity, education level, gender, gestational age of baby and insurance type. Additionally, our organization obtained data on neonatal intensive care level from the California Childcare services for all research performed by our groups. We also obtained hospital ownership type from Regional Perinatal Programs of California (RPPC).

2a2. RELIABILITY TESTING

***Note**: If accuracy/correctness (validity) of data elements was empirically tested, separate reliability testing of data elements is not required – in 2a2.1 check critical data elements; in 2a2.2 enter "see section 2b2 for validity testing of data elements"; and skip 2a2.3 and 2a2.4.*

2a2.1. What level of reliability testing was conducted? (*may be one or both levels*)

☐ **Critical data elements used in the measure** (*e.g., inter-abstractor reliability; data element reliability must address ALL critical data elements*)

☒ **Performance measure score** (e.g., *signal-to-noise analysis*)

2a2.2. For each level checked above, describe the method of reliability testing and what it tests (*describe the steps—do not just name a method; what type of error does it test; what statistical analysis was used*)

2011: This measure has been tested on California discharge data sets for several years (2004 to 2007) with ~560,000 births per year. We also examined these codes on the HCUP data set that comprised over 8 million births.

Reliability Testing was performed at the hospital level for 226 hospitals across California who had over 200 cases meeting our denominator inclusion criteria. For purposes of reliability testing, the 2011 California Statewide linked data was analyzed as described in the RAND Corporations "The Reliability of Provider Profiling: A Tutorial" by John L. Adams (RAND Corporation, TR-653-NCQA, 2009). This methodology is specifically recommended by the NQF to analyze the reliability of performance for performance measure scores.

Reliability in this context represents the ability of the proposed measure to confidently and accurately distinguish the performance of one entity (hospital) from another. As outlined in the RAND tutorial, "Conceptually, it is the ratio of signal to noise. The signal in this case is the proportion of variability in measured performance that can be explained by real differences in performance." There are 3 main drivers of reliability; sample size, differences between entities (hospitals), and measurement error.

Reliability is estimated using a beta-binomial model, which is appropriate for measuring the reliability of the UNC metric by hospital. The strategy involves fitting a beta-binomial model for the performance metric results. Two parameters (alpha and beta) that define the beta-binomial distribution are generated from the model. From these parameters, the "between hospital variance" was produced. Next, the within hospital variance was generated based on the proportion of affirmative answers. Analyzing the between hospital variance and the within hospital variance generates the reliability for each hospital site.

2a2.3. For each level of testing checked above, what were the statistical results from reliability testing? (e.*g., percent agreement and kappa for the critical data elements; distribution of reliability statistics from a signal-to-noise analysis*)
2011: There is concern for under-reporting of several of these diagnoses codes (especially those for hypoxia/asphyxia) which is why we went to the procedure codes for thoroughness of ascertainment. This is well supported by our earlier studies looking at Cesarean rates and neonatal outcomes (Gould, 2006).

The mean reliability score for the 226 hospitals was 0.91. Results are summarized by Hospital deciles in Table 1 below with the means, standard deviations, minimum and maximum reliability statistics and number of hospitals in each decile presented. Individual hospital performance score and reliability statistics are presented in Table 2.

Table 1: Unexpected Newborn Complications Hospital Results summarized by Hospital Deciles.

Reliability Statistics by Hospital Deciles					
Rank for Variable Reliability	Mean Reliability Statistic from Signal-to-Noise Analysis	Std Dev	N	Minimum	Maximum
1	0.71	0.08	22	0.55	0.81
2	0.84	0.02	23	0.82	0.87
3	0.88	0.01	23	0.87	0.90
4	0.91	0.00	22	0.90	0.92
5	0.93	0.01	23	0.92	0.94
6	0.94	0.01	23	0.94	0.95
7	0.96	0.00	22	0.95	0.96
8	0.97	0.00	23	0.96	0.97
9	0.97	0.00	23	0.97	0.98
10	0.98	0.00	22	0.98	0.99

Table 2: Unexpected Newborn Complications Hospital Results for each hospital used in the reliability testing

Hospital	Cases in Denominator	Cases in Numerator	Performance Score	Reliability Statistic from signal-to-noise analysis
1	1016	29	2.85	0.92
2	895	55	6.15	0.83
3	5837	220	3.77	0.98
4	1553	40	2.58	0.95
5	4389	48	1.09	0.99
6	2302	85	3.69	0.95
7	2870	106	3.69	0.96
8	234	9	3.85	0.67
9	291	6	2.06	0.82
10	731	19	2.60	0.90
11	3381	109	3.22	0.97
12	5330	169	3.17	0.98
13	917	23	2.51	0.92
14	1643	41	2.50	0.96
15	3592	126	3.51	0.97
16	3004	59	1.96	0.98
17	1256	30	2.39	0.94
18	1086	44	4.05	0.90
19	1977	43	2.18	0.97
20	2094	50	2.39	0.97
21	4984	116	2.33	0.99
22	1694	87	5.14	0.92
23	1311	33	2.52	0.94
24	4413	76	1.72	0.99
25	786	29	3.69	0.87
26	1058	56	5.29	0.87
27	680	31	4.56	0.83
28	2788	155	5.56	0.94
29	451	8	1.77	0.89
30	653	15	2.30	0.90
31	3069	61	1.99	0.98
32	685	49	7.15	0.77
33	1095	36	3.29	0.92
34	763	7	0.92	0.96
35	947	47	4.96	0.86
36	551	19	3.45	0.84
37	3621	110	3.04	0.97
38	1001	10	1.00	0.97
39	1149	55	4.79	0.89
40	1114	12	1.08	0.97
41	396	12	3.03	0.81
42	859	16	1.86	0.94

43	789	16	2.03	0.93
44	2616	188	7.19	0.93
45	648	20	3.09	0.87
46	1652	37	2.24	0.96
47	2882	51	1.77	0.98
48	452	12	2.65	0.85
49	1939	28	1.44	0.98
50	1720	57	3.31	0.94
51	2585	70	2.71	0.97
52	393	4	1.02	0.93
53	567	45	7.94	0.71
54	396	10	2.53	0.84
55	866	26	3.00	0.90
56	1009	33	3.27	0.91
57	319	23	7.21	0.60
58	5030	75	1.49	0.99
59	3291	172	5.23	0.95
60	3833	234	6.10	0.95
61	2506	77	3.07	0.96
62	2060	67	3.25	0.95
63	1988	40	2.01	0.97
64	991	19	1.92	0.94
65	1227	41	3.34	0.92
66	2829	69	2.44	0.97
67	1790	39	2.18	0.96
68	2406	31	1.29	0.98
69	3506	68	1.94	0.98
70	1693	79	4.67	0.92
71	2306	49	2.12	0.97
72	2736	98	3.58	0.96
73	3343	73	2.18	0.98
74	1317	51	3.87	0.92
75	1225	69	5.63	0.88
76	955	29	3.04	0.91
77	1890	55	2.91	0.95
78	2047	64	3.13	0.96
79	2782	167	6.00	0.94
80	2163	143	6.61	0.92
81	1810	104	5.75	0.91
82	1455	50	3.44	0.93
83	1119	51	4.56	0.89
84	4160	156	3.75	0.97
85	1638	88	5.37	0.91
86	2050	75	3.66	0.95
87	1933	75	3.88	0.94

88	1682	38	2.26	0.96
89	3416	88	2.58	0.98
90	3387	164	4.84	0.96
91	2845	126	4.43	0.96
92	957	57	5.96	0.84
93	1868	115	6.16	0.91
94	389	18	4.63	0.74
95	1379	29	2.10	0.95
96	603	90	14.93	0.60
97	1440	26	1.81	0.96
98	3134	301	9.60	0.92
99	403	18	4.47	0.75
100	1514	53	3.50	0.93
101	2367	43	1.82	0.98
102	995	70	7.04	0.83
103	482	17	3.53	0.82
104	650	20	3.08	0.87
105	1036	30	2.90	0.92
106	1734	65	3.75	0.94
107	2340	94	4.02	0.95
108	2567	104	4.05	0.95
109	830	38	4.58	0.86
110	1495	64	4.28	0.92
111	961	36	3.75	0.89
112	1945	59	3.03	0.95
113	1423	25	1.76	0.96
114	4154	178	4.29	0.97
115	1540	72	4.68	0.92
116	2521	79	3.13	0.96
117	764	28	3.66	0.87
118	1408	12	0.85	0.98
119	2322	57	2.45	0.97
120	1182	18	1.52	0.96
121	1747	85	4.87	0.92
122	2879	33	1.15	0.99
123	239	6	2.51	0.76
124	447	61	13.65	0.55
125	1393	37	2.66	0.94
126	368	15	4.08	0.75
127	1840	29	1.58	0.97
128	736	13	1.77	0.93
129	546	11	2.01	0.90
130	3089	142	4.60	0.96
131	986	25	2.54	0.93
132	1705	63	3.70	0.94

133	383	26	6.79	0.66
134	1579	24	1.52	0.97
135	1030	7	0.68	0.98
136	5495	384	6.99	0.96
137	2637	106	4.02	0.96
138	2217	61	2.75	0.96
139	613	12	1.96	0.91
140	2050	26	1.27	0.98
141	2085	33	1.58	0.98
142	1812	28	1.55	0.97
143	701	36	5.14	0.82
144	2481	143	5.76	0.94
145	2204	45	2.04	0.97
146	286	7	2.45	0.79
147	294	7	2.38	0.80
148	3242	63	1.94	0.98
149	1927	85	4.41	0.94
150	1638	28	1.71	0.97
151	3475	93	2.68	0.98
152	1655	28	1.69	0.97
153	4399	111	2.52	0.98
154	2033	74	3.64	0.95
155	503	11	2.19	0.88
156	1505	24	1.59	0.97
157	1892	123	6.50	0.91
158	379	9	2.37	0.84
159	889	135	15.19	0.69
160	1798	53	2.95	0.95
161	2266	102	4.50	0.94
162	1764	66	3.74	0.94
163	611	10	1.64	0.92
164	1926	48	2.49	0.96
165	3413	80	2.34	0.98
166	1092	25	2.29	0.94
167	381	13	3.41	0.79
168	809	30	3.71	0.88
169	1335	54	4.04	0.92
170	3453	150	4.34	0.96
171	1697	105	6.19	0.90
172	1231	76	6.17	0.87
173	494	12	2.43	0.87
174	2352	78	3.32	0.96
175	2857	148	5.18	0.95
176	6626	218	3.29	0.99
177	396	9	2.27	0.85

178	1574	57	3.62	0.93
179	1013	30	2.96	0.92
180	579	20	3.45	0.85
181	415	18	4.34	0.76
182	1297	28	2.16	0.95
183	2478	64	2.58	0.97
184	610	17	2.79	0.88
185	216	3	1.39	0.83
186	211	11	5.21	0.57
187	4033	112	2.78	0.98
188	2058	131	6.37	0.92
189	1779	38	2.14	0.96
190	600	10	1.67	0.92
191	2205	56	2.54	0.97
192	940	27	2.87	0.91
193	282	6	2.13	0.81
194	299	11	3.68	0.73
195	1130	33	2.92	0.93
196	809	33	4.08	0.87
197	243	14	5.76	0.59
198	878	53	6.04	0.83
199	1373	86	6.26	0.88
200	2465	96	3.89	0.95
201	4155	163	3.92	0.97
202	663	7	1.06	0.95
203	581	10	1.72	0.92
204	303	11	3.63	0.73
205	2779	68	2.45	0.97
206	2455	144	5.87	0.93
207	1036	13	1.25	0.96
208	659	19	2.88	0.88
209	1459	64	4.39	0.92
210	795	46	5.79	0.82
211	1478	124	8.39	0.86
212	1305	116	8.89	0.84
213	1641	66	4.02	0.93
214	722	14	1.94	0.92
215	1071	32	2.99	0.92
216	3447	68	1.97	0.98
217	586	15	2.56	0.88
218	1122	15	1.34	0.96
219	1617	122	7.54	0.88
220	1293	109	8.43	0.84
221	741	25	3.37	0.88
222	1568	20	1.28	0.98

223	1964	33	1.68	0.97
224	3271	86	2.63	0.98
225	1879	34	1.81	0.97
226	609	12	1.97	0.91

2a2.4 What is your interpretation of the results in terms of demonstrating reliability? (i.*e., what do the results mean and what are the norms for the test conducted?*)

Reliability scores can vary from 0.0 to 1.0, where a score of zero implies that all variation is attributable to measurement error (noise) whereas a reliability of 1.0 implies that the variation is caused by real differences in performance across hospitals. According to the RAND report, reliability scores of 0.7 are sufficient to see differences between hospitals and the mean, and scores around 0.9 are considered sufficient to see differences between pairs of hospitals.

The mean reliability score of our metric was 0.91 which is very good. Mean hospital scores in each of the 10 deciles were above 0.7, with 7 out of the 10 deciles having scores of over 0.91. The reliability testing results reveal that variation in scores is caused by real differences in performance across the hospitals and is not due to measurement error.

2b2. VALIDITY TESTING

2b2.1. What level of validity testing was conducted? (*may be one or both levels*)

☐ **Critical data elements** (*data element validity must address ALL critical data elements*)

☒ **Performance measure score**

- ☒ **Empirical validity testing**
- ☒ **Systematic assessment of face validity of <u>performance measure score</u> as an indicator** of quality or resource use (*i.e., is an accurate reflection of performance on quality or resource use and can distinguish good from poor performance*)

2b2.2. For each level of testing checked above, describe the method of validity testing and what it tests (*describe the steps—do not just name a method; what was tested, e.g., accuracy of data elements compared to authoritative source, relationship to another measure as expected; what statistical analysis was used)*

2011: Data correlations of the entire measure to other measures of have not been done. Comparisons of components codes to establish linkages to quality of care have been done by us and others (Gould, 1996; Gregory, 2009, among many others)

We performed validity testing at the level of the performance measure score. We will demonstrate both empirical validity testing as well as a systematic assessment of face validity of this measure as an indicator of quality.

<u>I: EMPIRICAL VALIDITY TESTING OF MEASURE SCORE:</u> Validity testing of this performance outcome measure was conducted by correlation and functional data analysis to demonstrate the relationship between our Unexpected Newborn Complications measure and admission to special care nursery, otherwise known as the "neonatal intensive care unit (NICU).

Neonatal admission to the NICU is a very good indication of an unexpected newborn complication in a <u>term infant with no preexisting conditions</u>. In the absence of a variable to indicate "admission to the NICU", the Unexpected Newborn Complications measure serves as a very good proxy in predicting which neonates will be admitted to the NICU. We predict that our measure will be highly correlated with a neonatal admission to the NICU. Unfortunately, a record of actual neonatal admission to the NICU is not available in the linked datasets that CMQCC has access to. As a result, we requested the National Perinatal Information Center (NPIC) to run a comparative analysis for us.

The National Perinatal Information Center (NPIC)'s Perinatal Center Data Base is comprised primarily of hospitals offering subspecialty perinatal services and includes academic and non-academic institutions. NPIC's member hospitals from across the country contribute approximately 725,000 perinatal discharges annually making it one of the largest repositories for hospital based perinatal clinical and financial discharge data in the United States. The large sample size contains a combination of clinical and financial information that enables users to perform linked analyses on maternal complications/comorbidities and infant outcomes.

NPIC uses the Unexpected Newborn Complications metric for performance and benchmarking in their member hospitals. One of the variables present in the NPIC dataset is "admission to special care nursery". This allows NPIC to compare the unexpected newborn complications measure to admissions to special care nursery in their member hospitals.

We requested NPIC to run an analyses on their 2011 data comparing unexpected newborn complications at their member hospitals to admissions to special care nursery (Neonatal intensive care unit) for newborns stratified by gestational age at birth.

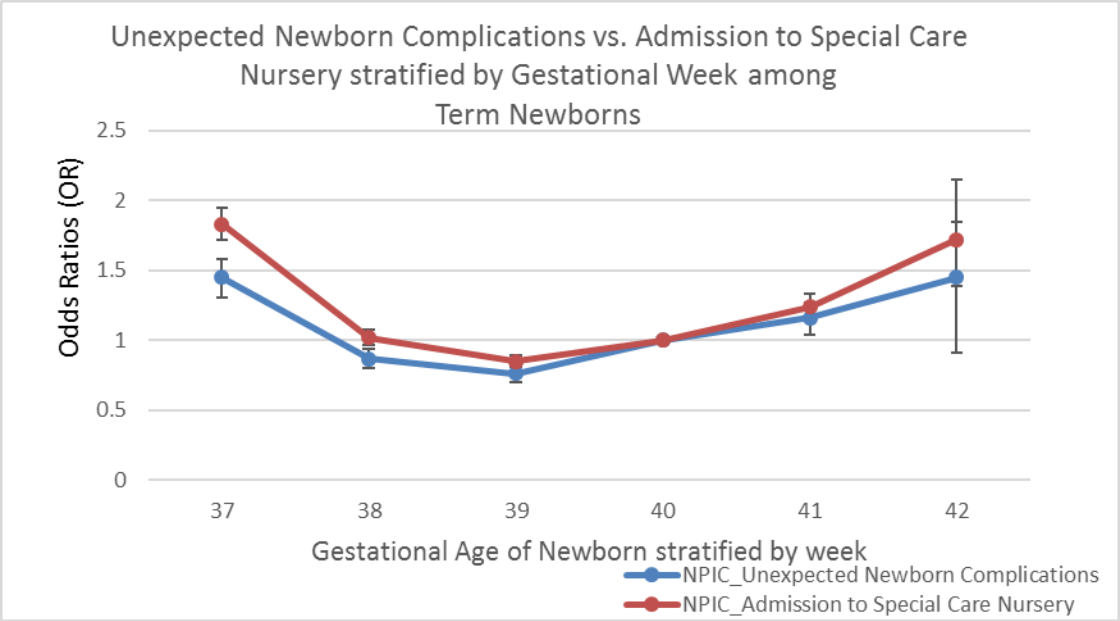
Listed in Table 4 are the odds ratios and confidence intervals for unexpected newborn complications vs admission to special care nursery stratified by gestational age. The neonatal population examined were term infants (37 weeks and 0 days to 42 weeks and 6 days). We compared the odds of total neonatal morbidity occurring at each gestational week to a reference week of 40 weeks. The odds ratios showed that neonatal morbidity in term babies followed a U-shaped curve, with the odds of experiencing morbidities in term babies being lowest at 39 and 40 weeks and the odds of experiencing morbidities being highest at 37 and 42 weeks. Similarly, the odds ratios showed that NICU admission in term babies also followed a U-shaped curve, with the odds of having a NICU admission being lowest at 39 and 40 weeks and the odds of having a NICU admission being highest at 37 and 42 weeks.

Table 4: Odds Ratio of admission to the Neonatal Intensive Care Unit at NPIC hospitals by gestational week compared to Unexpected Newborn Complications

Gestational Week	NPIC Unexpected Newborn Complications*	NPIC Admission to Special Care Nursery (NICU)
37	1.45 (1.32, 1.59)	1.83 (1.72, 1.95)
38	0.87 (0.80, 0.94)	1.02 (0.97, 1.08)
39	0.76 (0.71, 0.82)	0.85 (0.81, 0.89)
40	1	1
41	1.16 (1.05, 1.28)	1.24 (1.15, 1.33)
42	1.45 (1.05, 1.99)	1.72 (1.39, 2.15)

Note: The odds ratios for the Unexpected Newborn complications are unadjusted. The odds ratios describe the odds of an unexpected newborn complication by gestational week in the NPIC 2011 dataset.

Graph 1: Unexpected Newborn Complications vs. Admission to Special Care Nursery stratified by Gestational Week among Term Newborns.



We applied functional data analysis to compare both sets of U-shaped curves for neonatal morbidity and neonatal admission to the NICU using the adaptive Neyman test developed by Fan and Lin (1998). The observed curves in the first group (Unexpected Newborn complication results) are assumed to be a random sample from the model $X_j(t) = f_1(t) + \varepsilon_j(t)$ for t = 37,…,42 weeks and j = 1,…,6 where the random variables $\varepsilon_j(t)$ have a mean of zero. Similarly, the observed curves in the second group (neonatal intensive care unit admission results) are assumed to be a random sample from the model $Y_j(t) = f_2(t) + \varepsilon'_j(t)$ for t = 37,…,42 weeks and j = 1,…,6 where the random variables $\varepsilon'_j(t)$ have a mean of zero. We tested the hypothesis H0: f1(t) = f2(t) versus H1: f1(t) ≠ f2(t).

II: EMPIRICAL VALIDITY TESTING AS A BALANCING MEASURE

Use of Unexpected Newborn Complications as a balancing measure for QI projects that involve changing care during labor and birth. The most important outcomes for families are a healthy baby and a healthy mother. Therefore as we work on reducing cesarean births or 3rd/4th degree lacerations, we need to have a neonatal measure to show that there were no unintended consequences of the efforts, e.g. baby outcomes are as good (or better) than before the intervention began.

To test for utility as a balancing measure, we used the Unexpected Newborn Complications (UNC) measure in conjunction with 3 quality improvement projects focused on reduction of Nulliparous Term Singleton Vertex (NTSV) cesarean births (an NQF endorsed measure). The results of the first two pilots are shown in the figure (see Appendix 10) A summary of data for all 3 pilots is shown in the table below:

Hospital	Measure	Baseline: Mean for 8 quarters	QI Project: Mean for 6 quarters	%Change
Pilot 1	NTSV CS	32.5%	26.0%	Down 20%
	UNC	1.9%	1.8%	Down 5.0%
Pilot 2	NTSV CS	31.9%	26.0%	Down 18.5%
	UNC	6.2%	5.1%	Down 17.8%
Pilot 3	NTSV CS	27.8%	24.7%	Down 11.2%
	UNC	2.5%	1.9%	Down 24.0%

III: SYSTEMATIC ASSESSMENT OF FACE VALIDITY:

Face validity of the measure score as an indicator of quality was systematically assessed as follows. After the measure was fully specified, tested on multiple years of data, and its reliability was statistically assessed and determined to be very good, an expert group (other than those who advised on measure development) was assembled to rate face validity. The 22 experts included neonatologists, obstetricians and child care advocates (the list of experts is included in Appendix 9).

The expert group were provided with the detailed measure specifications and asked to rate their agreement with the following statement: "***This measure, as specified, provides an accurate reflection of quality and can be used to distinguish the quality of obstetric and neonatal care at the hospital level***."

The rating scale had 5 levels (1-5) with the following narrative anchors:

1=Strongly Disagree; 2= Somewhat Disagree; 3=Neutral; 4=Somewhat Agree; 5= Strongly Agree

2b2.3. What were the statistical results from validity testing? (*e.g., correlation; t-test*)

2011: Face validity was tested by discussions with both patient groups (see above) and with physician groups. THe later have long sought a way to measure healthy babies rather than just the more rare damaged infants. The inclusions were also seen by groups of obstetricians and neonatologists as fair and appropriate.

I: Empirical Validity Testing Results: We tested the hypothesis H0: $f1(t) = f2(t)$ versus H1: $f1(t) \neq f2(t)$. Applying the adaptive Neyman test to our data for the two U-shaped curves, we fail to reject the null hypothesis (P=0.13). Therefore, the data do not offer sufficient evidence to conclude that the two U-shaped curves are different.

In a separate test for correlation, the two sets of odds ratios exhibit high positive correlation: Pearson = 0.97 (0.77, 0.99) and Spearman = 0.93 (0.47, 0.99). If we test the hypothesis H0: true correlation is ≤ 0 versus H1: true correlation is > 0, then the data offer sufficient evidence to reject the null hypothesis. Therefore, we conclude that the true Pearson correlation is positive (P=0.014) and the true Spearman correlation is positive (P=0.0038).

II: Empirical Validity Testing as a Balancing Measure: In each Pilot hospital after successful intervention to reduce Nulliparous Singleton Term Vertex (NTSV) Cesarean births (down 15-22% from baseline), the Unexpected Newborn Complications (UNC) measure hospitals showed an actual decline in UNC after the QI initiative and none shown an increase. Hence we did NOT run formal statistics to see if UNC worsened. This demonstrates the utility of the Unexpected Newborn Complications measure as a balancing measure for NTSV Cesarean section.

III: Face Validity Results: The results of the face validity indicate that an independent group of experts (i.e.) different from those who advised on measure development) had high levels of agreement with the statement "***This measure, as specified, provides an accurate reflection of quality and can be used to distinguish the quality of obstetric and neonatal care at the hospital level***."

The mean rating was 4.82/5

Rating Scale	Number of experts who selected rating
1- Strongly Disagree	0
2- Somewhat Disagree	0
3-Neutral	1
4- Somewhat Agree	2

5- Strongly Agree	19
TOTAL	22

2b2.4. What is your interpretation of the results in terms of demonstrating validity? (i.*e., what do the results mean and what are the norms for the test conducted?*)

I: Empirical Validity Testing Results: The results from applying the adaptive Neyman test to the unexpected newborn complications odds ratios (stratified by gestational age) to a very similar measure (admissions to the neonatal intensive care unit in babies) in the NPIC cohort of babies demonstrate that the means of the odds ratios are very similar.

We can therefore conclude that severe and moderate morbidity in the NPIC cohort of newborns was validated by a very similar pattern of neonatal intensive care admissions with the lowest morbidity and NICU admissions being at 39 and 40 weeks and the highest at 37 and 42 weeks. This demonstrates that our metric successfully captures and quantifies neonatal morbidity in term newborns. Our results are also comparable to published studies on the differences in neonatal morbidity by gestational week (3-4) with the lowest adverse neonatal outcomes occurring at 39 and 40 weeks of gestation compared with later weeks of term gestation. Additionally, in a separate test of correlation, both sets of odds ratios demonstrated high positive correlation further reinforcing the results and conclusions discussed in the paragraph above.

II: Empirical Validity Testing as a Balancing Measure: In each Pilot hospital after successful intervention to reduce Nulliparous Singleton Term Vertex (NTSV) Cesarean births (down 15-22% from baseline), the Unexpected Newborn Complications measure was reduced, thus reassuring the medical staff that the intervention on the mother did not produce an adverse outcome in the baby. This demonstrates the utility of the Unexpected Newborn Complications measure as a balancing measure for NTSV Cesarean section.

III: Face Validity Interpretation

This measure was evaluated for face validity by a group of experts. Out of the 22 participants, 19 (86%) gave the measure a rating of 5, thereby strongly agreeing that the scores from the measure as specified would provide an accurate reflection of quality. The mean rating was very high 4.82/5 and none of the participants disagreed with the statement "***This measure, as specified, provides an accurate reflection of quality and can be used to distinguish the quality of obstetric and neonatal care at the hospital level"***

References:

1. Fan J. Test of Significance Based on Wavelet Thresholding and Neyman's Truncation. J of the American Statistical Assoc;1996; (91)674-88.
2. Fan J and Lin SK. Test of Significance When Data Are Curves. J of the American Statistical Assoc; 1998 (93)1007-21Zhang X and Kramer MS. Variations in Mortality and Morbidity by Gestational Age among Infants Born at Term. J Pediatr 2009;154:358-62
3. Tita AT, Landon MB, Spong CY et al. Timing of Elective Repeat Cesarean Delivery at Term and Neonatal Outcomes. N Engl J Med; 2009;360(2):111-120
4. Reddy UM, Bettegowda VR, Dias T et al. Term Pregnancy: A period of Heterogeneous Risk for Infant Mortality. Obstet Gynecol 2011;117(6):1279-1287

2b3. EXCLUSIONS ANALYSIS

NA ☐ no exclusions — *skip to section 2b4*

2b3.1. Describe the method of testing exclusions and what it tests (*describe the steps—do not just name a method; what was tested, e.g., whether exclusions affect overall performance scores; what statistical analysis was used*)

2011: The categories of exclusions will be discussed one by one:
Twins have very different issues than singletons and are often delivered for very different reasons and different times so they are hard to group together with healthy singletons. They comprise 1.5% of all births and only 0.5% of births over 37 weeks.
Preterm infants have very strong rates of morbidity and would overwhelm measures of term baby morbidity. Much of this is not preventable postnatally. There exists a NQF measure for administration of antenatal steroids for this population. Preterm births account for 12-13% of US births.
Congenital anomalies are an important source of neonatal morbidity but there is little to be done by the medical system for prevention. Many are prenatally diagnosed and so the family goes into the birth process knowing not to expect a "normal term newborn". Major congenital anomalies account for 1-2% of term births.
Likewise, Small for dates infants and infants with isoimmunization and drug withdrawal all have conditions acquired in utero and not in the birth process.

To identify singleton, hospital-born, babies born in California in 2011, we used a linked dataset of patient discharge data linked to vital statistics birth certificate records. Naval and military hospitals were not included as they do not submit patient discharge data to the Office of Statewide Health Planning and Development. Our analysis was limited to identifying morbidity occurring during the birth admission only. We did not track re-admissions.
After limiting our dataset to the above admissions, the following exclusions were analyzed for frequency and variability across hospitals included in our analysis:

- Low Birth Weight Babies (under 2500g)
- Preterm births (under 37) weeks and births above 47 weeks of gestation (to exclude nonsensical values of gestational age).
- Babies with Congenital Malformations (list of ICD-9 codes outlined in Appendix 2A)
- Babies with Pre-existing Conditions (list of ICD-9 codes outlined in Appendix 2B)
- Babies exposed to maternal drug use in-utero (list of ICD-9 codes outlined in Appendix 2C)

Note: Preterm births were excluded using the variable "Best Obstetric Estimate of Gestational Age" from the birth certificates. In the few cases that had missing values for Best Obstetric Estimate of Gestational Age, we used Gestational Age according to the Last Menstrual Period to exclude preterm births. If both sources of gestational age were missing, babies were required to have a birth-weight of 3000g for inclusion in the metric.

Exclusions were performed in a hierarchical manner in the order listed above. The exclusion steps described above are also detailed in the metric's algorithm flowcharts available on our website.

2b3.2. What were the statistical results from testing exclusions? (*include overall number and percentage of individuals excluded, frequency distribution of exclusions across measured entities, and impact on performance measure scores*)
We examined the overall frequencies and proportions of the admissions excluded for each exclusion criterion in the 2011 and 2012 California patient discharge-birth certificate linked data. After limiting the dataset to singleton, hospital-born babies, our initial cohort included 449,978 births in 2011 and 458,036 births in 2012. The final cohorts, after applying the additional exclusions described below included 392,328 births for 2011 and 397,864 births for 2012.
Results are presented below separately for 2011 and 2012. Categories are not mutually exclusive and statistical analyses present below were performed on the initial cohorts of singleton, hospital-born babies.

Table 5: 2011 data: (from among 449,978 birth admissions in 247 hospitals)

Exclusions	N	% Overall	Distribution across hospitals 25th, 50th, 75th percentile

Birth Weight Exclusion (<2500g) or over 8165g	22,875	5.08	3.42, 4.49, 5.50
Inappropriate Gestational age exclusion (<37 weeks or >47weeks) using Obstetrician's best estimate of Gestational Age	30,474	6.77	4.57, 6.0, 7.36
Inappropriate Gestational age exclusion GA_LMP Exclusion (if GA_OB_Estimate is missing, apply GA_LMP 37-47 wks)	429	0.095	0.041, 0.11, 0.27
Missing both GA and under 3000g Exclusion	107	0.024	0.0, 0.0, 0.0
Congenital Malformation Exclusion	14,795	3.33	1.4, 2.3, 3.5
Non-malformation Exclusion	12,904	2.87	1.45, 2.03, 3.19
Maternal Drug Use Exclusion	3,708	0.82	0.29, 0.62, 1.18

Table 6: 2012 data: (from among 458,036 births in 244 hospitals)

Exclusions	N	% Overall	Distribution across hospitals 25th, 50th, 75th percentile
Birth Weight Exclusion (<2500g) or over 8165g	23,323	5.09	3.38, 4.45, 5.49
Inappropriate Gestational age exclusion (<37 weeks or >47weeks)	31,234	6.82	4.66, 6.06, 7.40
Term GA_LMP Exclusion (if GA_OB_Estimate is missing, apply GA_LMP 37-47 wks)	550	0.14	0.047, 0.11, 0.38
Missing both GA and under 3000g Exclusion	153	0.03	0.0, 0.0, 0.0
Congenital Malformation Exclusion	15,078	3.79	1.59, 2.51, 3.59
Non-malformation Exclusion	14,074	3.07	1.34, 2.22, 3.39
Maternal Drug Use Exclusion	4,226	0.92	0.32, 0.69, 1.42

2b3.3. What is your interpretation of the results in terms of demonstrating that exclusions are needed to prevent unfair distortion of performance results? (*i.e., the value outweighs the burden of increased data collection and analysis. Note: **If patient preference is an exclusion**, the measure must be specified so that the effect on the performance score is transparent, e.g., scores with and without exclusion*)

The overall frequency of the exclusions is low (n=57,650, 12.8% for 2011) and (n=60,172, 13.1% for 2012). The distribution of exclusions across hospitals is as expected and random. However, we feel that each of the exclusions is absolutely necessary and should be retained as their inclusion would bias the performance results by including a subset of preterm, low-birth weight babies many of whom are already in poor health before birth. The inclusion of these children would bias the results and confound true performance differences in obstetric and neonatal quality at the hospital level. The metric's intended denominator population is healthy, full term neonates who are "expected to go home routinely" but unexpectedly experience adverse events.

Gestational week Exclusion:
Rationale: The measure excludes premature babies (those born before 37 weeks of gestation). In keeping with NCHS criteria for acceptable gestational ages, we used an upper range cut off of 46 weeks and 6 days for gestational age to exclude nonsensical values for gestational age.

Birth weight Exclusion:
Rationale: The measure excludes low birth weight infants as they may be premature, small for gestational age, or experienced intra-uterine growth restriction. Many low birth babies were also premature and would have been excluded anyway.

Congenital Malformation Exclusion:

Rationale: Babies with congenital malformations are excluded as they are generally not healthy and may have a myriad of conditions that require additional medical treatment soon after birth and later in life.

Pre-Existing Conditions Exclusion:
Rationale: Babies who are light for dates (small gestational age), experienced fetal growth retardation, who were affected by placenta previa, as well as fetuses affected by hemolytic disease due to Rh isoimmunization or hydrops were also excluded as they are not considered healthy and require additional medical treatment.

Maternal Drug Use Exclusion:
Rationale: Babies whose mothers used drugs during pregnancy were excluded as these infants would have suffered withdrawal, had longer neonatal lengths of stay and could have other health problems associated with exposure to drugs in-utero.

2b4. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES

If not an intermediate or health outcome, or PRO-PM, or resource use measure, skip to section 2b5.

2b4.1. What method of controlling for differences in case mix is used?

☒ **No risk adjustment or stratification**
☐ **Statistical risk model with** Click here to enter number of factors **risk factors**
☐ **Stratification by** Click here to enter number of categories **risk categories**
☐ **Other,** Click here to enter description

2b4.2. If an outcome or resource use measure is not risk adjusted or stratified, provide rationale and analyses to demonstrate that controlling for differences in patient characteristics (case mix) is not needed to achieve fair comparisons across measured entities.

2011: Risk adjustment is not done as with the exclusions above we feel that we have a homogenious enough population not to disadvantage a particular type of hospital. We do intend to test and potentially stratify hospitals by size as noted above.

RATIONALE
In the context of healthcare performance assessment, the purpose of a risk model is to reduce bias due to case mix characteristics present at the start of care (in this case the birth of the baby). This measure is not risk-adjusted. When constructing the measure, the exclusion criteria were chosen to ensure that the target population would be healthy, term babies with no pre-existing complications, thus reducing bias due to case mix complications. Babies more at risk for experiencing adverse outcomes (premature babies, low birth weight infants, babies with congenital malformations, exposure to maternal substance use and other pre-existing conditions) were excluded from the target population. The rationale for each of the exclusions is outlined in Question 2b 3.3.
The National Quality Forum prefers that measures are not risk adjusted for patient factors that could possibly obscure disparities (namely age, sex and socioeconomic status). As a result we did not adjust for sex or insurance status of the newborns. We chose not to adjust for gestational age as different morbidities are more prevalent at different gestational ages. Also, we did not want to mask morbidities resulting from early elective delivery practices (under 39 weeks of gestational age) or non-interventional practices in some hospitals (who do not induce women who are over 41 weeks pregnant, thus increasing the risk of stillbirth and morbidity in post term infants).
Variables related to quality of care are purposely not included in risk models for performance measures used to assess quality. Risk adjustment should not mask or adjust for the very factors that are driving the differences in neonatal health outcomes at hospitals across California. Accordingly, we did not adjust for a hospital's neonatal intensive care unit level, birth volume, ownership status, teaching status or number of maternal-fetal care specialists.

ANALYSES
We next considered maternal characteristics that might have an influence on neonatal outcomes. We examined the following variables:
-maternal age (under 20 years, 20-29 years, 30-34 years, 35 years and older)

-parity(nulliparous women vs. multiparous women)
-month prenatal care began (no prenatal care, first trimester, second trimester, third trimester)
-gestational diabetes in the mother
-chronic diabetes in the mother
-hypertension in the mother
ICD-9 Codes used for exploratory risk analysis are presented in Appendix 8
We also considered macrosomic babies (babies with a birth weight of over 4250g). Maternal age and prenatal care were not significant in the univariate analysis (p>0.0001) and were not included in the multivariate analysis. Logistic regression analysis was performed to assess the following risk factors (parity, month prenatal care began, gestational diabetes, chronic diabetes, hypertension and macrosomia). The outcome variable was dichotomous and was simply the occurrence of an adverse event (defined as the sum of either a severe or moderate morbidity) for each baby.

Results from Statistical Risk Adjustment

Statistic	
N	392,328
C-Statistic	0.61

Results from Hosmer and Lemeshow Statistic

Chi Square	**Degrees of Freedom**	**Probability**
10.59	2	0.0050

Expected probabilities based on the multivariate model were obtained and the adjusted metric for each hospital was calculated. Using methods outlined by Bailit and Grobman (2013) in their paper on Risk-adjustmed models for adverse obstetric outcomes ,we ranked hospitals and placed them in quintiles based on their unadjusted and adjusted metric value. When we compared adjusted and unadjusted metric values, ranks and quintiles for each hospital, we noticed very few hospitals (n=14, 5.7%) moved between quintiles. Statistically there was a high concordance (Kendall coefficient of concordance 0.94) between the unadjusted and adjusted ranks. However, there were hospitals where their rank based on their adjusted frequency differed substantially from their rank based on their unadjusted frequency (as much as 25 rank tiers).

In summary, while there are some individual factors that can statistically affect the score, when examined together at the hospital level they cancel each other out or are distributed evenly among hospitals so as not to significantly affect the rankings.

2b4.3. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors (clinical factors or sociodemographic factors) used in the statistical risk model or for stratification by risk (*e.g., potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of p<0.10; correlation of x or higher; patient factors should be present at the start of care*)
Not Applicable

2b4.4a. What were the statistical results of the analyses used to select risk factors?
Not Applicable

2b4.4b. Describe the analyses and interpretation resulting in the decision to select SDS factors (e.g. prevalence of the factor across measured entities, empirical association with the outcome, contribution of unique variation in the outcome, assessment of between-unit effects and within-unit effects)
Not Applicable
2b4.5. Describe the method of testing/analysis used to develop and validate the adequacy of the statistical model or stratification approach (*describe the steps—do not just name a method; what statistical analysis was used*)
Not Applicable

Provide the statistical results from testing the approach to controlling for differences in patient characteristics (case mix) below.
If stratified, skip to 2b4.9

2b4.6. Statistical Risk Model Discrimination Statistics (*e.g., c-statistic, R-squared*)**:**
Not Applicable
2b4.7. Statistical Risk Model Calibration Statistics (*e.g., Hosmer-Lemeshow statistic*):
Not Applicable
2b4.8. Statistical Risk Model Calibration – Risk decile plots or calibration curves:
Not Applicable
2b4.9. Results of Risk Stratification Analysis:
Not Applicable

2b4.10. What is your interpretation of the results in terms of demonstrating adequacy of controlling for differences in patient characteristics (case mix)? (i*.e., what do the results mean and what are the norms for the test conducted*)
Not Applicable

2b4.11. Optional Additional Testing for Risk Adjustment (*not required, but would provide additional support of adequacy of risk model, e.g., testing of risk model in another data set; sensitivity analysis for missing data; other methods that were assessed*)
Not Applicable

_____
2b5. IDENTIFICATION OF STATISTICALLY SIGNIFICANT & MEANINGFUL DIFFERENCES IN PERFORMANCE
2b5.1. Describe the method for determining if statistically significant and clinically/practically meaningful differences in performance measure scores among the measured entities can be identified (*describe the steps—do not just name a method; what statistical analysis was used? Do not just repeat the information provided related to performance gap in 1b)*
2011: Using California data (Patient Discharge Diagnosis sets) with >560,000 newborns reported each year, were identified that this measure falls into a reasonable bell shaped curve of hospital results. We have used both 95%tile at both ends and quintiles. It is not yet clear which will be superior for such needs as public reporting or benchmarking.

To examine differences in performance, we calculated 95% confidence intervals for the unadjusted metric results for all eligible hospitals. If a hospital's confidence interval did not include the California state mean (mean of the unexpected newborn complication results for all eligible hospitals in California) then the hospital was identified as statistically significantly better or worse than the California state average.

2b5.2. What were the statistical results from testing the ability to identify statistically significant and/or clinically/practically meaningful differences in performance measure scores across measured entities? (e.g., *number and percentage of entities with scores that were statistically significantly different from mean or some benchmark, different from expected; how was meaningful difference defined*)

The California state mean (the mean of all hospitals with eligible births in the state) was used as the benchmark. Using the approach described in 2b5.1, 53 hospitals (21.5%) of 247 California hospitals were rated as statistically significantly higher (worse) than the state mean and 78 hospitals (31.5%) were identified as statistically significantly lower (better) than the state mean. The other 116 hospitals were either higher or lower than the state mean but their results were not statistically significant. When all 247 hospitals were included in the calculation of the state mean, the summary statistics were as follows:

Mean	SD	Minimum	10th percentile	25th percentile	Median	75th percentile	90th percentile	Maximum

3.48	1.87	0.0	1.59	2.18	3.13	4.29	6.04	15.2

2b5.3. What is your interpretation of the results in terms of demonstrating the ability to identify statistically significant and/or clinically/practically meaningful differences in performance across measured entities? (i.*e., what do the results mean in terms of statistical and meaningful differences?*)
This measure is able to detect hospitals with better and worse than average performance. Hospitals that were identified as statistically significantly better or worse than the state average had scores that were at least 20% lower or 20% higher than the state mean, which we consider a meaningful difference in performance.

2b6. COMPARABILITY OF PERFORMANCE SCORES WHEN MORE THAN ONE SET OF SPECIFICATIONS
***If only one set of specifications, this section can be skipped*.**

Note: *This item is directed to measures that are risk-adjusted (with or without SDS factors)* ***OR*** *to measures with more than one set of specifications/instructions (e.g., one set of specifications for how to identify and compute the measure from medical record abstraction and a different set of specifications for claims or eMeasures). It does not apply to measures that use more than one source of data in one set of specifications/instructions (e.g., claims data to identify the denominator and medical record abstraction for the numerator).* ***Comparability is not required when comparing performance scores with and without SDS factors in the risk adjustment model. However, if comparability is not demonstrated for measures with more than one set of specifications/instructions, the different specifications (e.g., for medical records vs. claims) should be submitted as separate measures.***

2b6.1. Describe the method of testing conducted to compare performance scores for the same entities across the different data sources/specifications (*describe the steps—do not just name a method; what statistical analysis was used*)
Not Applicable

2b6.2. What were the statistical results from testing comparability of performance scores for the same entities when using different data sources/specifications? (*e.g., correlation, rank order*)
Not Applicable

2b6.3. What is your interpretation of the results in terms of the differences in performance measure scores for the same entities across the different data sources/specifications? (i.*e., what do the results mean and what are the norms for the test conducted*)
Not Applicable

2b7. MISSING DATA ANALYSIS AND MINIMIZING BIAS

2b7.1. Describe the method of testing conducted to identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and non-responders) and how the specified handling of missing data minimizes bias (*describe the steps—do not just name a method; what statistical analysis was used*)

A few variables are essential to the calculation of this metric. They are outlined below, and we have outlined safeguards to deal with the rare instances in which they are missing.

Missing V30.00 or V30.01 codes: These codes present in patient discharge data identify in-hospital singleton births. If these variables are missing, we are unable to identify in-hospital births and would have to exclude the case from the measure.

Missing Gestational Age (Using Best Obstetric Estimate of Gestational Age): In less than 1% of cases in linked administrative data, the best obstetric estimate of gestation age (from birth certificate data) is missing. In these cases, we use the LMP-based gestational age to identify full term infants. If both sources of gestational age are missing, we only include infants who are over 3000g, as they are more likely to be full term.

Missing Birth Weight: Birth weight is missing in less than 0.05% of cases in administrative data. If birth weight is missing, we have 2 sources of gestational age to be able to include an eligible baby into the metric.

Under-coding of diagnoses: The hierarchical construction of the numerator of this metric provides several double-checks to minimize the chance of missing a newborn with an unexpected newborn complication. If a truly sick newborn is missed at one stage of the metric, it can be captured in the next levels of the metric. We also require a length of stay for certain conditions to protect against over coding and under coding certain complications. Finally, if no complications are coded at all, a baby with a length of stay of over 5 days that does not have social reasons for remaining in hospital (adoption/foster care) or jaundice will be included in the metric. Of course, this measure will not identify newborn complications in infants born at home.

2b7.2. What is the overall frequency of missing data, the distribution of missing data across providers, and the results from testing related to missing data? (*e.g., results of sensitivity analysis of the effect of various rules for missing data/nonresponse; if no empirical sensitivity analysis, identify the approaches for handling missing data that were considered and pros and cons of each*)

This was answered in the question above (2b7.1) for the missing birth weight, gestational age and in-hospital birth variables. In the cases where diagnoses are under or over coded, **hospitals perform routine audits of hospital charts** if they find that their unexpected newborn complication rate is being driven by a particular diagnosis. Furthermore, we have worked with individual hospitals who have been able to identify coding practices to change resulting in improvement of their measure scores. One tertiary hospital changed its use of CPAP in keeping with regional norms, after it was found that CPAP was being mistakenly being over-coded in newborn records.

2b7.3. What is your interpretation of the results in terms of demonstrating that performance results are not biased due to systematic missing data (or differences between responders and non-responders) and how the specified handling of missing data minimizes bias**?** (i.*e., what do the results mean in terms of supporting the selected approach for missing data and what are the norms for the test conducted; if no empirical analysis, provide rationale for the selected approach for missing data*)

Missing data elements are too miniscule to have any effect on the measure. As stated in our responses to questions 2b7.1 and 2b7.2, the unexpected newborn complications measure has inbuilt checks to account for possible missing data, as well as under and over coding. It is highly unlikely that our measure will miss including an eligible infant.

3. Feasibility

Extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

3a. Byproduct of Care Processes

For clinical measures, the required data elements are routinely generated and used during care delivery (e.g., blood pressure, lab test, diagnosis, medication order).

3a.1. **Data Elements Generated as Byproduct of Care Processes.**

Coding/abstraction performed by someone other than person obtaining original information (E.g., DRG, ICD-9 codes on claims, chart abstraction for quality measure or registry)

If other:

3b. Electronic Sources

The required data elements are available in electronic health records or other electronic sources. If the required data are not in electronic health records or existing electronic sources, a credible, near-term path to electronic collection is specified.

3b.1. **To what extent are the specified data elements available electronically in defined fields?** *(i.e., data elements that are needed to compute the performance measure score are in defined, computer-readable fields)*

ALL data elements are in defined fields in a combination of electronic sources

3b.2. **If ALL the data elements needed to compute the performance measure score are not from electronic sources, specify a credible, near-term path to electronic capture, OR provide a rationale for using other than electronic sources.**

3b.3. **If this is an eMeasure, provide a summary of the feasibility assessment in an attached file or make available at a measure-specific URL.**

Attachment:

3c. Data Collection Strategy

Demonstration that the data collection strategy (e.g., source, timing, frequency, sampling, patient confidentiality, costs associated with fees/licensing of proprietary measures) can be implemented (e.g., already in operational use, or testing demonstrates that it is ready to put into operational use). For eMeasures, a feasibility assessment addresses the data elements and measure logic and demonstrates the eMeasure can be implemented or feasibility concerns can be adequately addressed.

3c.1. **Describe what you have learned/modified as a result of testing and/or operational use of the measure regarding data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, time and cost of data collection, other feasibility/implementation issues.**

IF a PRO-PM, consider implications for both individuals providing PROM data (patients, service recipients, respondents) and those whose performance is being measured.

1)Using a linked dataset: This measure was originally designed to be computable using Patient Discharge Data only. However, as the measure was put into production, it was surprisingly hard to identify term infants using Patient Discharge Diagnosis (PDD) data alone as called for in the initial specifications. Available ICD-9 codes for gestational age and birth weight are significantly under-utilized in practice. However, linking PDD to Birth Certificate data files has solved this issue as the Birth Certificate data fields for Best Obstetric Gestational Age and Birth weight have high degrees of completeness and accuracy. The Birth Certificate also allows the addition of extremely low Apgar scores as an additional numerator item.

(2)Coding Practices: Over and under coding: We learned that coding practices do vary for some ICD-9 codes with some hospitals being “over exuberant” in their coding and others clearly under-coding existing complications. The new specifications attempt to balance this issue by requiring that many codes for Moderate Complications additionally have an infant length of stay that exceeds the typical maternal postpartum length of stay (>2 days for a vaginal birth and >4 days for a cesarean birth). This requirement significantly reduces the number of infants identified but validates that these babies had significant morbidity. Conversely, some babies had very long neonatal length of stay without any codes to account for it, suggesting the possibility of under-coding. Our expert panel identified two categories of prolonged neonatal length of stay that were not medically serious and could be excluded from this consideration, namely neonatal jaundice typically treated with Bili-Lights, and social disruption for homelessness or foster care. We found that a number of babies with septicemia (771.81) had short length of stay indicating that it was not likely severe, therefore we added a requirement for a length of stay of at least 5 days to be included among Severe Complications.

3c.2. **Describe any fees, licensing, or other requirements to use any aspect of the measure as specified** *(e.g., value/code set, risk model, programming code, algorithm)*.
No fees or licensing required.

4. Usability and Use

Extent to which potential audiences (e.g., consumers, purchasers, providers, policy makers) are using or could use performance results for both accountability and performance improvement to achieve the goal of high-quality, efficient healthcare for individuals or populations.

4a. Accountability and Transparency

Performance results are used in at least one accountability application within three years after initial endorsement and are publicly reported within six years after initial endorsement (or the data on performance results are available). If not in use at the time of initial endorsement, then a credible plan for implementation within the specified timeframes is provided.

4.1. Current and Planned Use

NQF-endorsed measures are expected to be used in at least one accountability application within 3 years and publicly reported within 6 years of initial endorsement in addition to performance improvement.

Planned	Current Use (for current use provide URL)
Public Reporting Public Health/Disease Surveillance	Quality Improvement with Benchmarking (external benchmarking to multiple organizations) California Maternal Quality Care Collaborative https://www.cmqcc.org/focus-areas/quality-metrics/unexpected-complications-term-newborns National Perinatal Information Center http://www.npic.org/index.php Safe Deliveries Roadmap Collaborative: Washington State Hospital Association (WSHA) http://www.wsha.org/quality-safety/projects/safe-deliveries/ Oregon Perinatal Collaborative: Oregon Health Care Quality Corporation (QCorp) http://q-corp.org/maternity-care Quality Improvement (Internal to the specific organization) California Maternal Quality Care https://www.cmqcc.org/focus-areas/quality-metrics/unexpected-complications-term-newborns National Perinatal Information Center http://www.npic.org/index.php Safe Deliveries Roadmap Collaborative: Washington State Hospital Association (WSHA) http://www.wsha.org/quality-safety/projects/safe-deliveries/ Oregon Perinatal Collaborative: Oregon Health Care Quality Corporation (QCorp) http://q-corp.org/maternity-care

4a.1. For each CURRENT use, checked above, provide:

- Name of program and sponsor
- Purpose
- Geographic area and number and percentage of accountable entities and patients included

Current user 1: California Maternal Quality Care Collaborative
Purpose: Quality improvement with bench marking, internal quality improvement
Scope: All 250 California birthing hospitals representing just over 500,000 births annually.

Current User 2: National Perinatal Information Center(NPIC)
Purpose: Quality improvement with bench marking, internal quality improvement
Scope: All 82 NPIC member hospitals across the US representing 324,000 births annually

Current User 3:
Safe Deliveries Roadmap Collaborative: Washington State Hospital Association (WSHA)
Purpose: quality improvement with bench marking, internal quality improvement

Scope: 35 hospitals in Washington, Alaska and Montana representing approximately 175,000 newborns annually

Current User 4:
Oregon Perinatal Collaborative: Oregon Health Care Quality Corporation (QCorp)
Purpose: Quality improvement with bench marking, internal quality improvement
Scope: 14 hospitals in Oregon representing approximately 55,000 newborns annually.

4a.2. **If not currently publicly reported OR used in at least one other accountability application (e.g., payment program, certification, licensing) what are the reasons?** (*e.g., Do policies or actions of the developer/steward or accountable entities restrict access to performance results or impede implementation?*)
This measure was being tweaked in response to user feedback and we now feel that it is ready for public reporting. Our philosophy has been to provide hospitals with their own data for 2 to 3 years and then begin public reporting of hospital results. This measure has been shared state-wide on the California Maternal Data Center. Participating hospitals in California, Washington,Oregon, Alaska and Montana as well as 80+ National Perinatal Information Center (NPIC) member hospitals across the nation have access to this metric and have specific benchmarks and trends to follow their internal data. In addition, we provide extensive drill-down analysis for hospitals to understand why their rate may be elevated.
The Unexpected Newborn Complications metric specifications are available free of cost to any interested user on our CMQCC website.

4a.3. **If not currently publicly reported OR used in at least one other accountability application, provide a credible plan for implementation within the expected timeframes -- any accountability application within 3 years and publicly reported within 6 years of initial endorsement.** (*Credible plan includes the specific program, purpose, intended audience, and timeline for implementing the measure within the specified timeframes. A plan for accountability applications addresses mechanisms for data aggregation and reporting.*)
CMQCC provides hospital-level metrics for public reporting to the California Hospital Accountability and Reporting Task force (CHART), supported by the California Health Care Foundation. They and we are very interested in publicly reporting the Unexpected Newborn Complications measure especially as a balancing measure for Cesarean reduction efforts. We have committed to them that this measure will move to public reporting in California no later than after 2-3 years of confidential release. This will be with release of 2015 data later this year.

4b. Improvement

Progress toward achieving the goal of high-quality, efficient healthcare for individuals or populations is demonstrated. If not in use for performance improvement at the time of initial endorsement, then a credible rationale describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.

4b.1. **Progress on Improvement. (Not required for initial endorsement unless available.)**
Performance results on this measure (current and over time) should be provided in 1b.2 and 1b.4. Discuss:

- **Progress (trends in performance results, number and percentage of people receiving high-quality healthcare)**
- **Geographic area and number and percentage of accountable entities and patients included**

Since the initial endorsement in 2011, we have tweaked this measure in response to user feedback. At the time of initial endorsement, the measure was only in use in California hospitals. Since then hospitals in Washington, Oregon, Alaska and Montana hospitals are using it as well as National Perinatal Information Center (NPIC) member hospitals across the country (see 4a 1.).
In summary, we have:
a)Improved the measure since initial NQF endorsement in response to user (hospital) feedback.
b)Greatly widened the geographic area and number of states and hospitals across the country using the measure. Currently over 1 million births across the United States are evaluated using this measure. This corresponds to approximately 25% of all US births annually.
c)Allowed hospitals 2-3 years of internal data benchmarking before we provide hospital level results for public reporting to the California Hospital Accountability and Reporting Taskforce (CHART), supported by the California HealthCare Foundation.

4b.2. **If no improvement was demonstrated, what are the reasons? If not in use for performance improvement at the time of initial endorsement, provide a credible rationale that describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.**
We have not yet shown statewide improvement as the major use has been as a balancing measure. However, we have seen individual improvement in hospitals who have tried to directly improve their rates (10-20% reduction). We have not yet mounted a Quality Improvement collaborative for this measure but rather used it as a balancing measure for Cesarean Section Quality Improvement activities—i.e. we do not want to see an increase in poorer baby outcomes when the Cesarean Section rate goes

down. It has been critical in this endeavor as we were able to show no harm as we reduced Cesarean Section by 20% in a 3 hospital pilot project (60,000 babies over 2 years). Furthermore, we have worked with individual hospitals who have been able to identify practices to change that have improved their measure scores. One tertiary hospital identified that they were over-diagnosing and over-treating neonates for presumed sepsis at rates much higher than benchmarked tertiary facilities. Another hospital changed its use of CPAP in keeping with regional norms. Yet another hospital was able to show reduced birth injuries over time with standardized protocols for vacuum extractions

4c. Unintended Consequences

The benefits of the performance measure in facilitating progress toward achieving high-quality, efficient healthcare for individuals or populations outweigh evidence of unintended negative consequences to individuals or populations (if such evidence exists).

4c.1. **Were any unintended negative consequences to individuals or populations identified during testing; OR has evidence of unintended negative consequences to individuals or populations been reported since implementation? If so, identify the negative unintended consequences and describe how benefits outweigh them or actions taken to mitigate them.**

No unintended consequences have been noted. We are encouraging that this metric be used as a balancing measure to identify unintended consequences of other measures such as Cesarean Sections.

5. Comparison to Related or Competing Measures

If a measure meets the above criteria and there are endorsed or new related measures (either the same measure focus or the same target population) or competing measures (both the same measure focus and the same target population), the measures are compared to address harmonization and/or selection of the best measure.

5. Relation to Other NQF-endorsed Measures

Are there related measures (conceptually, either same measure focus or target population) or competing measures (conceptually both the same measure focus and same target population)? If yes, list the NQF # and title of all related and/or competing measures.

Yes

5.1a. **List of related or competing measures (selected from NQF-endorsed measures)**

5.1b. **If related or competing measures are not NQF endorsed please indicate measure title and steward.**

A formerly endorsed NQF measure (NQF # 0474 Birth Trauma -Injury to the Neonate) would have been considered a competing measure.

Measure Steward:Agency for Healthcare Research and Quality

5a. Harmonization

The measure specifications are harmonized with related measures;

OR

The differences in specifications are justified

5a.1. **If this measure conceptually addresses EITHER the same measure focus OR the same target population as NQF-endorsed measure(s):**

Are the measure specifications completely harmonized?

No

5a.2. **If the measure specifications are not completely harmonized, identify the differences, rationale, and impact on interpretability and data collection burden.**

There is no other currently endorsed measure in this topic area. A formerly endorsed NQF measure (NQF # 0474 Birth Trauma -Injury to the Neonate) would have been considered a "competing measure" as it conceptually addressed the same measure focus and target population. It suffered from over coding issues with several ICD codes dominating the measure that were ambiguous (e.g. "Other birth injuries NOS"). This remains an issue for ICD-10. For that and other reasons, that measure was "un-endorsed". Furthermore that measure was focused only on physical birth injuries while our measure identifies a much broader range of neonatal morbidities that are a consequence of labor and delivery.

5b. Competing Measures

The measure is superior to competing measures (e.g., is a more valid or efficient way to measure);

OR

Multiple measures are justified.

5b.1. **If this measure conceptually addresses both the same measure focus and the same target population as NQF-endorsed measure(s):**

Describe why this measure is superior to competing measures (e.g., a more valid or efficient way to measure quality); OR provide a rationale for the additive value of endorsing an additional measure. (Provide analyses when possible.)

NQF 0474 provides a very limited window into term morbidities from the single perspective of birth trauma. There are many other morbidities in term infants that are much more common and important to quantify as several of them are severe and can have long lasting implications well into childhood and beyond. We feel our measure is superior to NQF 0474 for the following reasons:

• We examine a much broader range of adverse events including deaths, transfers, low Apgar scores and a wide range of severe and moderate conditions. Including hypoxic encephalopathy, very low Apgar scores, and respiratory distress in term infants.

• After consulting neonatologists, pediatricians and obstetricians about the severity of certain conditions and how to quantify and group conditions appropriately, we are confident that our measure differentiates between severe and moderate morbidity.

• Our measure factors in neonatal length of stay, which is an important indicator in assessing whether an infant is truly severely ill or not. For example, an infant may have a diagnosis code for neonatal sepsis (a very serious newborn complication) but if the neonatal LOS was only 2 days (and no death or transfer) se We also include method of delivery and its impact on length of stay, as infants delivered via Cesarean section generally stay in hospital for four days and infants born vaginally stay for two days or less. We exclude

conditions like jaundice and social factors that cause infants to have longer neonatal lengths of stay.
• Our exclusions ensure that our denominator (target) population truly does consist of healthy term newborns by excluding preterm infants, low birth weight babies, congenital malformations, babies subjected to maternal drug use and other preexisting conditions.
• Our measure allows hospitals to drill down into sub-measures of morbidity such as respiratory complications, neurological complications and infections to determine what is driving their unexpected newborn complication rate.
•The larger incidence of conditions in our measure compared to the NQF birth injury measure allows for much better statistical analysis and discrimination. Furthermore, our measure is currently used to evaluate over 1 million births in multiple states and hospitals across the US (corresponding to approximately 25% of all US births).

Appendix

A.1 Supplemental materials may be provided in an appendix. All supplemental materials (such as data collection instrument or methodology reports) should be organized in one file with a table of contents or bookmarks. If material pertains to a specific submission form number, that should be indicated. Requested information should be provided in the submission form and required attachments. There is no guarantee that supplemental materials will be reviewed.
Attachment **Attachment:** Unexpected_Newborn_Complications_Appendices-635908793270671228.xlsx

Contact Information

Co.1 **Measure Steward (Intellectual Property Owner):** California Maternal Quality Care Collaborative
Co.2 **Point of Contact:** Elliott, Main, main@cmqcc.org, 415-992-2252-
Co.3 **Measure Developer if different from Measure Steward:** California Maternal Quality Care Collaborative
Co.4 **Point of Contact:** Elliott, Main, main@cmqcc.org, 415-992-2252-

Additional Information

Ad.1 **Workgroup/Expert Panel involved in measure development**
Provide a list of sponsoring organizations and workgroup/panel members' names and organizations. Describe the members' role in measure development.
CMQCC members: Elliott Main, MD; Anisha Abreo MPH, Debra Bingham RN DrPH; Kathryn Melsop, MS
CPQCC members: Terri Slagle, MD and Richard Powers, MD (both Neonatologists long active in QI research)
MQI members: Kimberly Gregory, M.D., MPH; Lisa Korst, MD PhD; Moshe Freedman, PhD; Sonal Shah, MPH; Michael Lu, MD MPH.
The entire team reviewed and discussed the concepts and ICD9 codes. MQI did the first pass of the data analysis, CMQCC did subsequent. Testing with focus groups and with other organizations was done by CMQCC.

Measure Developer/Steward Updates and Ongoing Maintenance
Ad.2 **Year the measure was first released:** 2009
Ad.3 **Month and Year of most recent revision:** 07, 2013
Ad.4 **What is your frequency for review/update of this measure?** Every 2-3 years
Ad.5 **When is the next scheduled review/update for this measure?** 12, 2017

Ad.6 **Copyright statement:** This will be in the public domain.
Ad.7 **Disclaimers:**

Ad.8 **Additional Information/Comments:** In addition to the ICD-9 code version of the Unexpected Newborn Complications Specs that is located at the URL provided in S.1, we have also provided an ICD-10 version of this metric located in the attached Appendices.

MEASURE WORKSHEET

This document summarizes the evaluation of the measure as it progresses through NQF's Consensus Development Process (CDP). The information submitted by measure developers/stewards is included after the Brief Measure Information, Preliminary Analysis, and Pre-meeting Public and Member Comments sections.

To navigate the links in the worksheet: Ctrl + click link to go to the link; ALT + LEFT ARROW to return

Brief Measure Information

NQF #: 1382
Measure Title: Percentage of low birthweight births
Measure Steward: Centers for Disease Control and Prevention
Brief Description of Measure: The percentage of births with birthweight <2,500 grams
Developer Rationale: The percentage of low birthweight infants has increased by 22% from 6.7% of births in 1984 to 8.2% in 2007. Since a substantially lower percentage of low birthweight births has already been achieved in the United States in the past, there appears to be no reason why a substantially lower level could not be achieved again.

Numerator Statement: The number of babies born weighing <2,500 grams at birth in the study population
Denominator Statement: All births in the study population
Denominator Exclusions: None

Measure Type: Outcome
Data Source: Patient Reported Data/Survey
Level of Analysis: Population : County or City, Population : National, Population : Regional

IF Endorsement Maintenance – Original Endorsement Date: Aug 15, 2011 **Most Recent Endorsement Date:** Aug 15, 2011

Maintenance of Endorsement -- Preliminary Analysis

To maintain NQF endorsement endorsed measures are evaluated periodically to ensure that the measures still meets the NQF endorsement criteria ("maintenance"). The emphasis for maintaining endorsement is focused on how effective the measure is for promoting improvements in quality. Endorsed measures should have some experience from the field to inform the evaluation. The emphasis for maintaining endorsement is noted for each criterion.

Criteria 1: Importance to Measure and Report

1a. Evidence
Maintenance measures – less emphasis on evidence unless there is new information or change in evidence since the prior evaluation.

1a. Evidence. The evidence requirements for a health outcomes measure include providing rationale that supports the relationship of the health outcome to processes or structures of care. The guidance for evaluating the clinical evidence asks if the relationship between the measured health outcome and at least one clinical action is identified and supported by the stated rationale.

Summary of evidence:

- This is a population health measure that has been collected and reported for many years.
- In 2006, the infant mortality rate for low birthweight infants was 55.38 infant deaths per 1,000 live births, 25 times the rate of 2.24 for infants born weighing 2,500 grams or more.
- Prenatal care can assist women in eliminating or successfully managing pregnancy risk factors such as smoking during pregnancy, inadequate weight gain, pregnancy-associated diabetes, and others. Women who resolve pregnancy risks can substantially lower their chance of having a low birthweight infant. (Source: Ricketts SA,

Murray EK, Schwalberg R. Reducing low birthweight by resolving risks: results from Colorado´s prenatal plus program. Am J Public Health. 2005 Nov;95(11):1952-7. Epub 2005 Sep 29)

Changes to evidence from last review

☒ **The developer attests that there have been no changes in the evidence since the measure was last evaluated.**

☐ **The developer provided updated evidence for this measure:**

Question for the Committee:

- *This is a population measure for states and the nation. Are there actions within the healthcare systems that can (or have) affected the results?*

Preliminary rating for evidence: ☒ Pass ☐ No Pass

1b. Gap in Care/Opportunity for Improvement and 1b. Disparities
Maintenance measures – increased emphasis on gap and variation

1b. Performance Gap. The performance gap requirements include demonstrating quality problems and opportunity for improvement.

- The US percentage of low birthweight births increased by 22% from 1984 until a 2006 high of 8.26%. The US percentage of low birthweight births is substantially higher than in most other developed countries.
- Recent vital statistics data have shown a decline in percentage of low birthweight to 8.02% in 2013 and 8.00% in 2014.

Disparities

- 2014 birth data shows variation in low birthweight for various race/ethnicity:
  White – 7% Black – 12.8% Hispanic 7.1%
  American Indian.Alaskan native – 8.1% Asian/Pacific Islander 8.1%

Questions for the Committee:

- *This is a population measure for states and the nation. Does this measure provide a meaningful context and goal for the collective healthcare efforts in perinatal care?*
- *Is the decrease in low birthweight since 2006 a reflection of improved healthcare for moms and babies?*
- *Do you expect the rates of low birthweight to continue to decline?*
- *Is there a gap in care that warrants a national performance measure?*

Preliminary rating for opportunity for improvement: ☒ High ☐ Moderate ☐ Low ☐ Insufficient

Committee pre-evaluation comments

Criteria 1: Importance to Measure and Report (including 1a, 1b, 1c)

1a. Evidence to Support Measure Focus

Comments:

Yes.

1b. Performance Gap

Comments:

Yes, there appears to be a gap between various race/ethnicity.

1c. Composite Performance Measure

Comments:

** Data elements are clearly defined and the logic is clear.**

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability

2a1. Reliability Specifications

Maintenance measures – no change in emphasis – specifications should be evaluated the same as with new measures

2a1. Specifications requires the measure, as specified, to produce consistent (reliable) and credible (valid) results about the quality of care when implemented.

Data source(s): Birth records/vital statistics

Specifications:

- The levels of analysis are geographic populations such as counties/cities, regions/states and the nation.
- Numerator is number of births weighing less than 2500 grams.
- Denominator is all live births.
- There are no exclusions or risk-adjustment.

Questions for the Committee:

- *Are all the data elements clearly defined? Are all appropriate codes included?*
- *Is the logic or calculation algorithm clear?*
- *Is it likely this measure can be consistently implemented?*

2a2. Reliability Testing Testing attachment

Maintenance measures – less emphasis if no new testing data provided

2a2. Reliability testing demonstrates if the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise enough to distinguish differences in performance across providers.

For maintenance measures, summarize the reliability testing from the prior review:

Data element validity was assessed by direct comparison of birth certificate data to medical records. Sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) were computed. Data element validity tested against a "gold standard" such as the medical record also counts for reliability.

Describe any updates to testing None submitted

SUMMARY OF TESTING

Reliability testing level ☐ **Measure score** ☒ **Data element** ☐ **Both**

Reliability testing performed with the data source and level of analysis indicated for this measure ☐ **Yes** ☐ **No**

Method(s) of reliability testing Data element validity testing – see validity section

Results of reliability testing

Guidance from the Reliability Algorithm

Precise specifications (Box 1) → empiric reliability testing g(Box 2) → empiric testing of patient level data → use rating from validity testing of patient-level data elements

Questions for the Committee:

- *Since no new testing data is provided does the Committee agree the criterion has already been met and further discussion and voting is not needed?*

Preliminary rating for reliability: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

2b. Validity
Maintenance measures – less emphasis if no new testing data provided

2b1. Validity: Specifications

2b1. Validity Specifications. This section should determine if the measure specifications are consistent with the evidence.

Specifications consistent with evidence in 1a. ☒ **Yes** ☐ **Somewhat** ☐ **No**
Specification not completely consistent with evidence

Question for the Committee:

- *Are the specifications consistent with the evidence?*

2b2. Validity testing

2b2. Validity Testing should demonstrate the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality.

For maintenance measures, summarize the validity testing from the prior review:
Data element validity testing as described below

Describe any updates to validity testing none

SUMMARY OF TESTING
Validity testing level ☐ **Measure score** ☒ **Data element testing against a gold standard** ☐ **Both**

Method of validity testing of the measure score:

- ☐ **Face validity only**
- ☐ **Empirical validity testing of the measure score**

Validity testing method:
Data element validity testing as described below

Validity testing results:

- Study 1 – Roohan PJ, Josberger RE, Acar J et al. Validation of birth certificate data in New York State. Journal of Community Health 2003;28:335-46.
  - Study 1 – Low birthweight (<2,500 g) - Sensitivity, specificity, PPV, NPV all 100%.
- Study 2 – DiGuiseppe DL, Aron DC, Ranbom L et al. Reliability of birth certificate data: A multi-hospital comparison to medical records information. Maternal and Child Health Journal 2002;6:169-179.
  - Study 2 – Birthweight <3000 g or > 3000 g. Concordance 99%, sensitivity 99% specificity 99% PPV 100% NPV 98%.

2b3-2b7. Threats to Validity

2b3. Exclusions: NA

-

2b4. Risk adjustment: **Risk-adjustment method** ☒ **None** ☐ **Statistical model** ☐ **Stratification**

2b5. Meaningful difference *(can statistically significant and clinically/practically meaningful differences in performance measure scores can be identified)*:

- Data are based on the complete population of 4.2 million birth certificates filed in the United States

each year.

- Any statistically significant increase or decrease, using standard methods for significance testing.

Question for the Committee:

- *Does this measure identify meaningful differences about quality?*

2b6. Comparability of data sources/methods: NA

2b7. Missing Data not addressed

Guidance from the algorithm

Aligned with specifications (Box 1) → Threats to validity addressed (Box 2) →empirical validity testing (Box 3) →testing with patient-level data elements (Box 10) →appropriate method of testing (Box 11) → high/moderate certainty the data are valid → moderate

Preliminary rating for validity: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Question for the Committee

- ***As no new testing has been provided, does the Committee agree that the criterion has been met and no further discussion or voting is warranted?***

Committee pre-evaluation comments

Criteria 2: Scientific Acceptability of Measure Properties (including all 2a, 2b, and 2d)

2a1. & 2b1. Specifications

Comments:

No new data has been added.

2a2. Reliability Testing

Comments:

2b2. Validity Testing

Comments:

2b3. Exclusions Analysis

2b4. Risk Adjustment/Stratification for Outcome or Resource Use Measures

2b5. Identification of Statistically Significant & Meaningful Differences In Performance

2b6. Comparability of Performance Scores When More Than One Set of Specifications

2b7. Missing Data Analysis and Minimizing Bias

Comments:

Criterion 3. Feasibility

Maintenance measures – no change in emphasis – implementation issues may be more prominent

3. Feasibility is the extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

- Birth certificate data is an established data source
- The data are in electronic form

Questions for the Committee:

- *Are the required data elements routinely generated and used during care delivery?*
- *Are the required data elements available in electronic form, e.g., EHR or other electronic sources?*
- *Is the data collection strategy ready to be put into operational use?*

Preliminary rating for feasibility: ☒ **High** ☐ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 3: Feasibility

3a. Byproduct of Care Processes
3b. Electronic Sources
3c. Data Collection Strategy
Comments:
Should be electronically available.

Criterion 4: Usability and Use
Maintenance measures – increased emphasis – much greater focus on measure use and usefulness, including both impact /improvement and unintended consequences

4. Usability and Use evaluate the extent to which audiences (e.g., consumers, purchasers, providers, policymakers) use or could use performance results for both accountability and performance improvement activities.

Current uses of the measure
Publicly reported? ☒ **Yes** ☐ **No**

Current use in an accountability program? ☐ **Yes** ☐ **No**
OR
Planned use in an accountability program? ☐ **Yes** ☐ **No**

Accountability program details
Public reporting by CDC National Vital Statistics System http://www.cdc.gov/nchs/births.htm

Improvement results From the peak of 8.26% in 2006 the percentage of low birthweight has decline to 8.00% in 2014.

Unexpected findings (positive or negative) during implementation None identified

Potential harms

Feedback:

Questions for the Committee:

- *Is this measure information useful and meaningful to patients, payers, policymakers?*
- *Does the information from this measure reflect the collective efforts of the healthcare system in delivery perinatal care?*
- *How can the performance results be used to further the goal of high-quality, efficient healthcare?*
- *Do the benefits of the measure outweigh any potential unintended consequences?*

Preliminary rating for usability and use: ☒ **High** ☐ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 4: Usability and Use

4a. Accountability and Transparency
4b. Improvement
4c. Unintended Consequences
Comments:

Criterion 5: Related and Competing Measures

Related or competing measures
None

Harmonization
NA

Pre-meeting public and member comments

•

1. IMPACT, OPPORTUITY, EVIDENCE - IMPORTANCE TO MEASURE AND REPORT

Importance to Measure and Report is a threshold criterion that must be met in order to recommend a measure for endorsement. All three subcriteria must be met to pass this criterion. See guidance on evidence.

Measures must be judged to be important to measure and report in order to be evaluated against the remaining criteria. (evaluation criteria)

1c.1 **Structure-Process-Outcome Relationship** *(Briefly state the measure focus, e.g., health outcome, intermediate clinical outcome, process, structure; then identify the appropriate links, e.g., structure-process-health outcome; process- health outcome; intermediate clinical outcome-health outcome)*:

In 2006, the infant mortality rate for low birthweight infants was 55.38 infant deaths per 1,000 live births, 25 times the rate of 2.24 for infants born weighing 2,500 grams or more. For very low birthweight infants (<1,500 grams), the infant mortality rate was 240.44 infant deaths per 1,000 live births, 107 times the rate for normal birthweight infants. Source: Mathews T.J., MacDorman MF. Infant mortality statistics from the 2006 period linked birth/infant death data set. National vital statistics reports vol 58 no 17. Hyattsville, MD: April 2010.

1c.2-3 **Type of Evidence** *(Check all that apply)*:

Other

Linked birth and infant death certificate data for the entire US population

1c.4 **Directness of Evidence to the Specified Measure** *(State the central topic, population, and outcomes addressed in the body of evidence and identify any differences from the measure focus and measure target population)*:

Prenatal care can assist women in eliminating or successfully managing pregnancy risk factors such as smoking during pregnancy, inadequate weight gain, pregnancy-associated diabetes, and others. Women who resolve pregnancy risks can substantially lower their chance of having a low birthweight infant.

Source: Ricketts SA, Murray EK, Schwalberg R. Reducing low birthweight by resolving risks: results from Colorado´s prenatal plus program. Am J Public Health. 2005 Nov;95(11):1952-7. Epub 2005 Sep 29.

1c.5 **Quantity of Studies in the Body of Evidence** *(Total number of studies, not articles)*:

1c.6 **Quality of Body of Evidence** (*Summarize the certainty or confidence in the estimates of benefits and harms to patients across studies in the body of evidence resulting from study factors. Please address: a) study design/flaws; b) directness/indirectness of the evidence to this measure (e.g., interventions, comparisons, outcomes assessed, population included in the evidence); and c) imprecision/wide confidence intervals due to few patients or events):*

1c.7 **Consistency of Results across Studies** *(Summarize the consistency of the magnitude and direction of the effect):*

1c.8 **Net Benefit** *(Provide estimates of effect for benefit/outcome; identify harms addressed and estimates of effect; and net benefit - benefit over harms)*:

1c.9 **Grading of Strength/Quality of the Body of Evidence**. Has the body of evidence been graded?

1c.10 **If body of evidence graded, identify the entity that graded the evidence including balance of representation and any disclosures regarding bias:**

1c.11 **System Used for Grading the Body of Evidence:** To my knowledge, the evidence has not been formally rated, but since it is based on an accurate population data source, and these relationships have been found each year for the past 30 years of data collection, I believe that they constitute high quality evidence.

1c.12 **If other, identify and describe the grading scale with definitions:**

1c.13 **Grade Assigned to the Body of Evidence:** Women who smoke during pregnancy and who have late or no prenatal care have a higher percentage of low birthweight births, and higher infant mortality rates. This is from national birth certificate data and these

relationships have been stable in the data each year since we began measuring these variables.

1c.14 Summary of Controversy/Contradictory Evidence: None

1c.15 Citations for Evidence other than Guidelines*(Guidelines addressed below)*:
Martin JA, Hamilton BE, Sutton PD et al. Births: Final data for 2007. National vital statistics reports, vol 58 no 24, August 2010.

1c.16 Quote verbatim, the specific guideline recommendation *(Including guideline # and/or page #)*:
Healthy People 2010 Objective 16-10: Reduce low birthweight and very low birthweight.

1c.17 Clinical Practice Guideline Citation:
http://www.healthypeople.gov/hpscripts/KeywordResult.asp?n269=269&n362=362&Submit=Submit

1c.18 National Guideline Clearinghouse or other URL:
http://www.healthypeople.gov/hpscripts/KeywordResult.asp?n269=269&n362=362&Submit=Submit

1c.19 Grading of Strength of Guideline Recommendation. Has the recommendation been graded?

1c.20 If guideline recommendation graded, identify the entity that graded the evidence including balance of representation and any disclosures regarding bias:

1c.21 System Used for Grading the Strength of Guideline Recommendation:

1c.22 If other, identify and describe the grading scale with definitions:

1c.23 Grade Assigned to the Recommendation: N/A

1c.24 Rationale for Using this Guideline Over Others: Scientific acceptability. Widely-used measure. Easy to measure, use and understand.

Based on the NQF descriptions for rating the evidence, what was the developer's assessment of the quantity, quality, and consistency of the body of evidence?
1c.25 Quantity: **1c.26** Quality: **1c.27** Consistency:

1. Evidence, Performance Gap, Priority – Importance to Measure and Report

Extent to which the specific measure focus is evidence-based, important to making significant gains in healthcare quality, and improving health outcomes for a specific high-priority (high-impact) aspect of healthcare where there is variation in or overall less-than-optimal performance. ***Measures must be judged to meet all subcriteria to pass this criterion and be evaluated against the remaining criteria.***

1a. Evidence to Support the Measure Focus – See attached Evidence Submission Form
1382_Evidence_MSF5.0_Data.doc

1b. Performance Gap

Demonstration of quality problems and opportunity for improvement, i.e., data demonstrating:

- considerable variation, or overall less-than-optimal performance, in the quality of care across providers; and/or
- disparities in care across population groups.

1b.1. **Briefly explain the rationale for this measure** *(e.g., the benefits or improvements in quality envisioned by use of this measure)*
The percentage of low birthweight infants has increased by 22% from 6.7% of births in 1984 to 8.2% in 2007. Since a substantially lower percentage of low birthweight births has already been achieved in the United States in the past, there appears to be no reason why a substantially lower level could not be achieved again.

1b.2. **Provide performance scores on the measure as specified (current and over time) at the specified level of analysis.** *(This is required for endorsement maintenance. Include mean, std dev, min, max, interquartile range, scores by decile. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included). This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*
The US percentage of low birthweight births has increased by 22% since 1984. The US percentage of low birthweight births is substantially higher than in most other developed countries.

1b.3. **If no or limited performance data on the measure as specified is reported in 1b2, then provide a summary of data from the literature that indicates opportunity for improvement or overall less than optimal performance on the specific focus of measurement.**
1. Martin JA, Hamilton BE, Sutton PD et al. Births: Final data for 2007. National vital statistics reports, vol 58 no 24, August 2010.

2. Organization for Economic Cooperation and Development Health Data 2010. Available at: http://www.ecosante.org/index2.php?base=OCDE&langh=ENG&langs=ENG&sessionid=

1b.4. **Provide disparities data from the measure as specified (current and over time) by population group, e.g., by race/ethnicity, gender, age, insurance status, socioeconomic status, and/or disability.** *(This is required for endorsement maintenance. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities include.) This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*
In 2007, the percentage of low birthweight births was 13.9% for non-Hispanic black women, 1.9 times the 7.3% for non-Hispanic white women. The higher percentage of low birthweight infants for non-Hispanic black women accounts for much of their elevated infant mortality risk, when compared to non-Hispanic white women.

1b.5. **If no or limited data on disparities from the measure as specified is reported in 1b4, then provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement. Include citations.**
Martin JA, Hamilton BE, Sutton PD et al. Births: Final data for 2007. National vital statistics reports, vol 58 no 24, August 2010.

1c. High Priority (previously referred to as High Impact)

The measure addresses:

- a specific national health goal/priority identified by DHHS or the National Priorities Partnership convened by NQF; OR
- a demonstrated high-priority (high-impact) aspect of healthcare (e.g., affects large numbers of patients and/or has a substantial impact for a smaller population; leading cause of morbidity/mortality; high resource use (current and/or future); severity of illness; and severity of patient/societal consequences of poor quality).

1c.1. **Demonstrated high priority aspect of healthcare**
Affects large numbers, A leading cause of morbidity/mortality, High resource use, Severity of illness, Patient/societal consequences of poor quality

1c.2. **If Other:**

1c.3. **Provide epidemiologic or resource use data that demonstrates the measure addresses a high priority aspect of healthcare. List citations in 1c.4.**
Here is a quotation from reference 1 below:
"Infants born at low birth weight (LBW) – conventionally defined as a birth weight less than 2,500 grams – experience severe health and developmental difficulties that can impose substantial costs on society. For example, the expected costs of delivery and initial care of a baby weighing 1000 grams at
birth can exceed $100,000 (in year 2000 dollars), and the risk of death within one year of birth is over one-in-five. Even among babies weighing 2000-2100 grams, who have comparatively low mortality rates, an additional pound (454 grams) of weight is still associated with a $10,000 difference in hospital
charges for inpatient services."

1c.4. **Citations for data demonstrating high priority provided in 1a.3**
1. Almond D, Chay KY, Lee DS. The costs of low birthweight. National Bureau of Economic Research, Working Paper 10552, June 2004. Available at:
http://www.nber.org/papers/w10552.

2. Petrou S, Eddama O, Mangham L. Arch Dis Child Fetal Neonatal Ed. 2010 May 20. [Epub ahead of print] A structured review of the recent literature on the economic consequences of preterm birth.

3. Dorling J, D'Amore A, Salt A, et al. Data collection from very low birthweight infants in a geographical region: methods, costs, and trends in mortality, admission rates, and resource utilisation over a fi ve-year period. Early Hum Dev 2006;82:117–24.

4. Tommiska V, Tuominen R, Fellman V. Economic costs of care in extremely
low birthweight infants during the fi rst 2 years of life. Pediatr Crit Care Med 2003;4:157–63.

5. Russell RB, Green NS, Steiner CA, et al. Cost of hospitalization for preterm and low birth weight infants in the United States. Pediatrics 2007;120:e1–9.
6. Mistry H, Dowie R, Franklin RC, Jani BR. Acta Paediatr. 2009 Jul;98(7):1123-9. Epub 2009 Apr 30.Costs of neonatal care for low-birthweight babies in English hospitals.

1c.5. **If a PRO-PM (e.g. HRQoL/functional status, symptom/burden, experience with care, health-related behaviors), provide evidence that the target population values the measured PRO and finds it meaningful.** *(Describe how and from whom their input was obtained.)*

2. Reliability and Validity—Scientific Acceptability of Measure Properties

Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. ***Measures must be judged to meet the subcriteria for both reliability and validity to pass this criterion and be evaluated against the remaining criteria.***

2a.1. Specifications The measure is well defined and precisely specified so it can be implemented consistently within and across organizations and allows for comparability. eMeasures should be specified in the Health Quality Measures Format (HQMF) and the Quality Data Model (QDM).

De.5. **Subject/Topic Area** *(check all the areas that apply):*
Perinatal and Reproductive Health, Perinatal and Reproductive Health : Perinatal

De.6. **Cross Cutting Areas** *(check all the areas that apply):*

S.1. **Measure-specific Web Page** *(Provide a URL link to a web page specific for this measure that contains current detailed specifications including code lists, risk model details, and supplemental materials. Do not enter a URL linking to a home page or to general information.)*
ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/Dataset_Documentation/DVS/natality/UserGuide2007.pdf

S.2a. **If this is an eMeasure**, HQMF specifications must be attached. Attach the zipped output from the eMeasure authoring tool (MAT) - if the MAT was not used, contact staff. (Use the specification fields in this online form for the plain-language description of the specifications)
Attachment:

S.2b. **Data Dictionary, Code Table, or Value Sets** *(and risk model codes and coefficients when applicable) must be attached. (Excel or csv file in the suggested format preferred - if not, contact staff)*
URL **Attachment:**

S.3. **For endorsement maintenance**, please briefly describe any changes to the measure specifications since last endorsement date and explain the reasons.

S.4. **Numerator Statement** *(Brief, narrative description of the measure focus or what is being measured about the target population, i.e., cases from the target population with the target process, condition, event, or outcome)*
IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
The number of babies born weighing <2,500 grams at birth in the study population

S.5. **Time Period for Data** *(What is the time period in which data will be aggregated for the measure, e.g., 12 mo, 3 years, look back to August for flu vaccination? Note if there are different time periods for the numerator and denominator.)*
A calendar year (for example, 2010)

S.6. **Numerator Details** *(All information required to identify and calculate the cases from the target population with the target process, condition, event, or outcome such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
IF an OUTCOME MEASURE, describe how the observed outcome is identified/counted. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
Data are directly available from public-use data files of national birth certificate data produced by the National Center for Health Statistics.

S.7. **Denominator Statement** *(Brief, narrative description of the target population being measured)*
All births in the study population

S.8. **Target Population Category** *(Check all the populations for which the measure is specified and tested if any):*
Children's Health, Maternal Health

S.9. **Denominator Details** *(All information required to identify and calculate the target population/denominator such as definitions, specific data collection items/responses , code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
Data are directly available from public-use data files of national birth certificate data produced by the National Center for Health Statistics.

S.10. **Denominator Exclusions** *(Brief narrative description of exclusions from the target population)*
None

S.11. **Denominator Exclusion Details** *(All information required to identify and calculate exclusions from the denominator such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
None

S.12. **Stratification Details/Variables** *(All information required to stratify the measure results including the stratification variables, definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format with at S.2b)*
- Stratify the measure by single vs. multiple births
- Stratify the measure by birthweight of less thant 1,500 grams (i.e. very low birthweight) vs. 1,500-2,499 grams (i.e. moderately low birthweight).

S.13. **Risk Adjustment Type** (Select type. Provide specifications for risk stratification in S.12 and for statistical model in S.14-15)
No risk adjustment or risk stratification
If other:

S.14. **Identify the statistical risk model method and variables** *(Name the statistical method - e.g., logistic regression and list all the risk factor variables. Note - risk model development and testing should be addressed with measure testing under Scientific Acceptability)*
N/A

S.15. **Detailed risk model specifications** *(must be in attached data dictionary/code list Excel or csv file. Also indicate if available at measure-specific URL identified in S.1.)*
Note: Risk model details (including coefficients, equations, codes with descriptors, definitions), should be provided on a separate worksheet in the suggested format in the Excel or csv file with data dictionary/code lists at S.2b.

S.15a. **Detailed risk model specifications** (*if not provided in excel or csv file at S.2b*)

S.16. **Type of score:**
Other
If other: Percentage

S.17. **Interpretation of Score** *(Classifies interpretation of score according to whether better quality is associated with a higher score, a lower score, a score falling within a defined interval, or a passing score)*
Better quality = Lower score

S.18. **Calculation Algorithm/Measure Logic** *(Describe the calculation of the measure score as an ordered sequence of steps including identifying the target population; exclusions; cases meeting the target process, condition, event, or outcome; aggregating data; risk adjustment; etc.)*
The number of births weighing <2,500 grams/Total births at any birthweight * 100

S.19. **Calculation Algorithm/Measure Logic Diagram URL or Attachment** *(You also may provide a diagram of the Calculation Algorithm/Measure Logic described above at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*

S.20. **Sampling** *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*

IF a PRO-PM, identify whether (and how) proxy responses are allowed.
This measure is based on the complete population of 4.2 million births in the United States each year. As such, it is not a sample and is not subject to sampling limitations.

S.21. **Survey/Patient-reported data** *(If measure is based on a survey, provide instructions for conducting the survey and guidance on minimum response rate.)*
IF a PRO-PM, specify calculation of response rates to be reported with performance measure results.

S.22. **Missing data** (specify how missing data are handled, e.g., imputation, delete case.)
Required for Composites and PRO-PMs.

S.23. **Data Source** *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*
If other, please describe in S.24.
Patient Reported Data/Survey

S.24. **Data Source or Collection Instrument** *(Identify the specific data source/data collection instrument e.g. name of database, clinical registry, collection instrument, etc.)*
IF a PRO-PM, identify the specific PROM(s); and standard methods, modes, and languages of administration.
National Center for Health Statistics, Natality Detail file. These publicly available data files contain individual record data for the 4.2 million births in the United States each year. Data are from birth certificates.

S.25. **Data Source or Collection Instrument** *(available at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
URL

S.26. **Level of Analysis** *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*
Population : County or City, Population : National, Population : Regional

S.27. **Care Setting** *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*
Hospital/Acute Care Facility, Other
If other: United States, states, counties

S.28. **COMPOSITE Performance Measure** - Additional Specifications *(Use this section as needed for aggregation and weighting rules, or calculation of individual performance measures if not individually endorsed.)*

2a. Reliability – See attached Measure Testing Submission Form
2b. Validity – See attached Measure Testing Submission Form
1382_MeasureTesting_MSF5.0_Data.doc

2. RELIABILITY & VALIDITY - SCIENTIFIC ACCEPTABILITY OF MEASURE PROPERTIES

Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. **(evaluation criteria)**
Measure testing must demonstrate adequate reliability and validity in order to be recommended for endorsement. Testing may be conducted for data elements and/or the computed measure score. Testing information and results should be entered in the appropriate field. Supplemental materials may be referenced or attached in item 2.1. See guidance on measure testing.

2a2. Reliability Testing. (*Reliability testing was conducted with appropriate method, scope, and adequate demonstration of reliability.*)

2a2.1 Data/Sample *(Description of the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)***:**
Many studies have found a high degree of reliability in the percent low birthweight measure from the birth certificate. I describe two examples below.

Study 1. 110 birth certificates were randomly sampled from each of 4 different counties in New York State. Total sample size = 440. Birth certificates were traced back to their hospital of origin and birth certificate data were directly compared to hospital medical record data.

Study 2. A random sample of birth certificates from 20 hospitals in the Cleveland metropolitan area. Total sample size =33,616

2a2.2 Analytic Method *(Describe method of reliability testing & rationale)***:**
Study 1 - Direct comparison of birth certificate data to medical records. Sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) were computed.

Study 2 – Direct comparison of birth certificate data to data from medical records collected by the Cleveland Health Quality Choice Initiative, a voluntary regional initiative to compare hospital performance. Concordance, sensitivity, specificity, PPV, and NPV were computed.

2a2.3 Testing Results *(Reliability statistics, assessment of adequacy in the context of norms for the test conducted)***:**
Study 1 – Low birthweight (<2,500 g) - Sensitivity, specificity, PPV, NPV all 100%.

Study 2 – Birthweight <3000 g or > 3000 g. Concordance 99%, sensitivity 99% specificity 99% PPV 100% NPV 98%.

Source: Study 1 – Roohan PJ, Josberger RE, Acar J et al. Validation of birth certificate data in New York State. Journal of Community Health 2003;28:335-46.

Study 2 – DiGuiseppe DL, Aron DC, Ranbom L et al. Reliability of birth certificate data: A multi-hospital comparison to medical records information. Maternal and Child Health Journal 2002;6:169-179.

2b. VALIDITY. Validity, Testing, including all Threats to Validity: H☐ M☐ L☐ I ☐

2b1.1 Describe how the measure specifications *(measure focus, target population, and exclusions)* **are consistent with the evidence cited in support of the measure focus (***criterion 1c)* **and identify any differences from the evidence:**

2b2. Validity Testing. (*Validity testing was conducted with appropriate method, scope, and adequate demonstration of validity.)*

2b2.1 Data/Sample *(Description of the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)***:**
Many studies have found a high degree of validity in the percent low birthweight measure from the birth certificate. I describe two examples below.

Study 1. 110 birth certificates were randomly sampled from each of 4 different counties in New York State. Total sample size = 440. Birth certificates were traced back to their hospital of origin and birth certificate data were directly compared to hospital medical record data.

Study 2. A random sample of birth certificates from 20 hospitals in the Cleveland metropolitan area. Total sample size =33,616

2b2.2 Analytic Method *(Describe method of validity testing and rationale; if face validity, describe systematic assessment)*:
Study 1 - Direct comparison of birth certificate data to medical records. Sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) were computed.

Study 2 – Direct comparison of birth certificate data to data from medical records collected by the Cleveland Health Quality Choice Initiative, a voluntary regional initiative to compare hospital performance. Concordance, sensitivity, specificity, PPV, and NPV were computed.

2b2.3 Testing Results *(Statistical results, assessment of adequacy in the context of norms for the test conducted; if face validity, describe results of systematic assessment)*:
Study 1 – Low birthweight (<2,500 g) - Sensitivity, specificity, PPV, NPV all 100%.

Study 2 – Birthweight <3000 g or > 3000 g. Concordance 99%, sensitivity 99% specificity 99% PPV 100% NPV 98%.

Source: Study 1 – Roohan PJ, Josberger RE, Acar J et al. Validation of birth certificate data in New York State. Journal of Community Health 2003;28:335-46.

Study 2 – DiGuiseppe DL, Aron DC, Ranbom L et al. Reliability of birth certificate data: A multi-hospital comparison to medical records information. Maternal and Child Health Journal 2002;6:169-179.

POTENTIAL THREATS TO VALIDITY. (*All potential threats to validity were appropriately tested with adequate results.*)

2b3. Measure Exclusions. (*Exclusions were supported by the clinical evidence in 1c or appropriately tested with results demonstrating the need to specify them.*)

2b3.1 Data/Sample for analysis of exclusions *(Description of the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)*:

2b3.2 Analytic Method (*Describe type of analysis and rationale for examining exclusions, including exclusion related to patient preference)*:

2b3.3 Results *(Provide statistical results for analysis of exclusions, e.g., frequency, variability, sensitivity analyses):*

2b4. Risk Adjustment Strategy. (*For outcome measures, adjustment for differences in case mix (severity) across measured entities was appropriately tested with adequate results.*)

2b4.1 Data/Sample *(Description of the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)*:
No risk adjustment needed.

2b4.2 Analytic Method (*Describe methods and rationale for development and testing of risk model or risk stratification including selection of factors/variables)*:

2b4.3 Testing Results *(Statistical risk model: Provide quantitative assessment of relative contribution of model risk factors; risk model performance metrics including cross-validation discrimination and calibration statistics, calibration curve and risk decile plot, and assessment of adequacy in the context of norms for risk models. Risk stratification: Provide quantitative assessment of relationship of risk factors to the outcome and differences in outcomes among the strata)*:

2b4.4 If outcome or resource use measure is not risk adjusted, provide rationale and analyses to justify lack of adjustment:

2b5. Identification of Meaningful Differences in Performance. (*The performance measure scores were appropriately analyzed and discriminated meaningful differences in quality.*)

2b5.1 Data/Sample *(Describe the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)*:
Data are based on the complete population of 4.2 million birth certificates filed in the United States each year.

2b5.2 Analytic Method *(Describe methods and rationale to identify statistically significant and practically/meaningfully differences in performance)*:
Any statisically significant increase or decrease,using standard methods for significance testing.

2b5.3 Results *(Provide measure performance results/scores, e.g., distribution by quartile, mean, median, SD, etc.; identification of statistically significant and meaningfully differences in performance)*:
N/A - no scores needed

2b6. Comparability of Multiple Data Sources/Methods. (*If specified for more than one data source, the various approaches result in comparable scores.*)

2b6.1 Data/Sample *(Describe the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)*:
Multiple data sources and/or methods are not needed as birth certificate data provide the gold standard for any measurement of this variable.

2b6.2 Analytic Method *(Describe methods and rationale for testing comparability of scores produced by the different data sources specified in the measure)*:
N/A

2b6.3 Testing Results *(Provide statistical results, e.g., correlation statistics, comparison of rankings; assessment of adequacy in the context of norms for the test conducted)*:
N/A

2c. Disparities in Care: H☐ M☐ L☐ I ☐ NA☐ (*If applicable, the measure specifications allow identification of disparities.*)

2c.1 If measure is stratified for disparities, provide stratified results *(Scores by stratified categories/cohorts)*: Number and percent of low, very low, and moderately low birthweight births by
single vs. multiple birth: United States, 2008

Birthweight (grams)	Number of births			% low birthweight		
	Total	Single	Twin/+	Total	Single	Twin/+
Total	4247694	4102766	144928			
2499 or less	347209	262479	84730	8.18	6.40	58.62
1499 or less	61773	45441	16332	1.46	1.11	11.30
1500 - 2499	285436	217038	68398	6.73	5.30	47.32
2500 or more	3896124	3836320	59804			
Not stated	4361	3967	394			

2c.2 If disparities have been reported/identified (e.g., in 1b), but measure is not specified to detect disparities, please explain:
N/A

2.1-2.3 Supplemental Testing Methodology Information:

Steering Committee: Overall, was the criterion, *Scientific Acceptability of Measure Properties*, met?
(*Reliability and Validity must be rated moderate or high*) Yes☐ No☐
Provide rationale based on specific subcriteria:

If the Committee votes No, STOP

3. Feasibility

Extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

3a. Byproduct of Care Processes

For clinical measures, the required data elements are routinely generated and used during care delivery (e.g., blood pressure, lab test, diagnosis, medication order).

3a.1. **Data Elements Generated as Byproduct of Care Processes.**

Other

If other: Data are from birth certificates filed for each US birth.

3b. Electronic Sources

The required data elements are available in electronic health records or other electronic sources. If the required data are not in electronic health records or existing electronic sources, a credible, near-term path to electronic collection is specified.

3b.1. **To what extent are the specified data elements available electronically in defined fields?** *(i.e., data elements that are needed to compute the performance measure score are in defined, computer-readable fields)*

Yes

3b.2. **If ALL the data elements needed to compute the performance measure score are not from electronic sources, specify a credible, near-term path to electronic capture, OR provide a rationale for using other than electronic sources.**

3b.3. **If this is an eMeasure, provide a summary of the feasibility assessment in an attached file or make available at a measure-specific URL.**

Attachment:

3c. Data Collection Strategy

Demonstration that the data collection strategy (e.g., source, timing, frequency, sampling, patient confidentiality, costs associated with fees/licensing of proprietary measures) can be implemented (e.g., already in operational use, or testing demonstrates that it is ready to put into operational use). For eMeasures, a feasibility assessment addresses the data elements and measure logic and demonstrates the eMeasure can be implemented or feasibility concerns can be adequately addressed.

3c.1. **Describe what you have learned/modified as a result of testing and/or operational use of the measure regarding data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, time and cost of data collection, other feasibility/implementation issues.**

IF a PRO-PM, consider implications for both individuals providing PROM data (patients, service recipients, respondents) and those whose performance is being measured.

Data are of high quality and no modifications are needed.

3c.2. **Describe any fees, licensing, or other requirements to use any aspect of the measure as specified** *(e.g., value/code set, risk model, programming code, algorithm)*.

4. Usability and Use

Extent to which potential audiences (e.g., consumers, purchasers, providers, policy makers) are using or could use performance results for both accountability and performance improvement to achieve the goal of high-quality, efficient healthcare for individuals or populations.

4a. Accountability and Transparency

Performance results are used in at least one accountability application within three years after initial endorsement and are publicly reported within six years after initial endorsement (or the data on performance results are available). If not in use at the time of initial endorsement, then a credible plan for implementation within the specified timeframes is provided.

4.1. Current and Planned Use

NQF-endorsed measures are expected to be used in at least one accountability application within 3 years and publicly reported within 6 years of initial endorsement in addition to performance improvement.

Planned	Current Use (for current use provide URL)
Public Reporting Quality Improvement (Internal to the specific organization)	

4a.1. For each CURRENT use, checked above, provide:

- Name of program and sponsor
- Purpose
- Geographic area and number and percentage of accountable entities and patients included

4a.2. If not currently publicly reported OR used in at least one other accountability application (e.g., payment program, certification, licensing) what are the reasons? (*e.g., Do policies or actions of the developer/steward or accountable entities restrict access to performance results or impede implementation?*)

4a.3. If not currently publicly reported OR used in at least one other accountability application, provide a credible plan for implementation within the expected timeframes -- any accountability application within 3 years and publicly reported within 6 years of initial endorsement. (*Credible plan includes the specific program, purpose, intended audience, and timeline for implementing the measure within the specified timeframes. A plan for accountability applications addresses mechanisms for data aggregation and reporting.*)

4b. Improvement

Progress toward achieving the goal of high-quality, efficient healthcare for individuals or populations is demonstrated. If not in use for performance improvement at the time of initial endorsement, then a credible rationale describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.

4b.1. Progress on Improvement. (Not required for initial endorsement unless available.)

Performance results on this measure (current and over time) should be provided in 1b.2 and 1b.4. Discuss:

- **Progress (trends in performance results, number and percentage of people receiving high-quality healthcare)**
- **Geographic area and number and percentage of accountable entities and patients included**

4b.2. If no improvement was demonstrated, what are the reasons? If not in use for performance improvement at the time of initial endorsement, provide a credible rationale that describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.

4c. Unintended Consequences

The benefits of the performance measure in facilitating progress toward achieving high-quality, efficient healthcare for individuals or populations outweigh evidence of unintended negative consequences to individuals or populations (if such evidence exists).

4c.1. **Were any unintended negative consequences to individuals or populations identified during testing; OR has evidence of unintended negative consequences to individuals or populations been reported since implementation? If so, identify the negative unintended consequences and describe how benefits outweigh them or actions taken to mitigate them.**

N/A - no data problems have been identified or are expected.

5. Comparison to Related or Competing Measures

If a measure meets the above criteria and there are endorsed or new related measures (either the same measure focus or the same target population) or competing measures (both the same measure focus and the same target population), the measures are compared to address harmonization and/or selection of the best measure.

5. Relation to Other NQF-endorsed Measures
Are there related measures (conceptually, either same measure focus or target population) or competing measures (conceptually both the same measure focus and same target population)? If yes, list the NQF # and title of all related and/or competing measures.

5.1a. **List of related or competing measures (selected from NQF-endorsed measures)**

5.1b. **If related or competing measures are not NQF endorsed please indicate measure title and steward.**

5a. Harmonization

The measure specifications are harmonized with related measures;
OR
The differences in specifications are justified

5a.1. **If this measure conceptually addresses EITHER the same measure focus OR the same target population as NQF-endorsed measure(s):**
Are the measure specifications completely harmonized?

5a.2. **If the measure specifications are not completely harmonized, identify the differences, rationale, and impact on interpretability and data collection burden.**

5b. Competing Measures

The measure is superior to competing measures (e.g., is a more valid or efficient way to measure);
OR
Multiple measures are justified.

5b.1. **If this measure conceptually addresses both the same measure focus and the same target population as NQF-endorsed measure(s):**
Describe why this measure is superior to competing measures (e.g., a more valid or efficient way to measure quality); OR provide a rationale for the additive value of endorsing an additional measure. (Provide analyses when possible.)

Appendix

A.1 Supplemental materials may be provided in an appendix. All supplemental materials (such as data collection instrument or methodology reports) should be organized in one file with a table of contents or bookmarks. If material pertains to a specific submission form number, that should be indicated. Requested information should be provided in the submission form and required attachments. There is no guarantee that supplemental materials will be reviewed.

Attachment:

Contact Information

Co.1 **Measure Steward (Intellectual Property Owner):** Centers for Disease Control and Prevention

Co.2 **Point of Contact:** Joyce, Martin, jcm9@cdc.gov, 301-458-4362-
Co.3 **Measure Developer if different from Measure Steward:** Centers for Disease Control and Prevention
Co.4 **Point of Contact:** Joyce, Martin, jcm9@cdc.gov, 301-458-4362-

Additional Information

Ad.1 **Workgroup/Expert Panel involved in measure development**
Provide a list of sponsoring organizations and workgroup/panel members' names and organizations. Describe the members' role in measure development.

Measure Developer/Steward Updates and Ongoing Maintenance
Ad.2 **Year the measure was first released:**
Ad.3 **Month and Year of most recent revision:**
Ad.4 **What is your frequency for review/update of this measure?** This measure has been used in vital statistics data since the 1930´s. Data quality reviewed annually
Ad.5 **When is the next scheduled review/update for this measure?**

Ad.6 **Copyright statement:**
Ad.7 **Disclaimers:**

Ad.8 **Additional Information/Comments:**

MEASURE WORKSHEET

This document summarizes the evaluation of the measure as it progresses through NQF's Consensus Development Process (CDP). The information submitted by measure developers/stewards is included after the Brief Measure Information, Preliminary Analysis, and Pre-meeting Public and Member Comments sections.

To navigate the links in the worksheet: Ctrl + click link to go to the link; ALT + LEFT ARROW to return

Brief Measure Information

NQF #: 1391
Measure Title: Frequency of Ongoing Prenatal Care (FPC)
Measure Steward: National Committee for Quality Assurance
Brief Description of Measure: The percentage of Medicaid deliveries that had the following number of expected prenatal visits:

- less than 21 percent of expected visits.
- 21 percent–40 percent of expected visits.
- 41 percent–60 percent of expected visits.
- 61 percent–80 percent of expected visits.
- greater than or equal to 81 percent of expected visits.

Developer Rationale: This measure assesses appropriate prenatal care by seeking to ensure that pregnant women receive the proper number of prenatal care visits for each stage of a pregnancy. Women who utilize prenatal care can minimize their risk for pregnancy complications and negative birth outcomes, which include higher risk of infant morbidity and mortality. Adherence to this measure could improve access to services and provide opportunities to improve health of both the infant and the mother.

Numerator Statement: Women who had the appropriate number of expected prenatal visits
Denominator Statement: The percentage of deliveries of live births between November 6 of the year prior to the measurement year and November 5 of the measurement year.
Denominator Exclusions: Exclude non-live births

Measure Type: Process
Data Source: Administrative claims, Electronic Clinical Data, Paper Medical Records
Level of Analysis: Health Plan, Integrated Delivery System

IF Endorsement Maintenance – Original Endorsement Date: Aug 15, 2011 **Most Recent Endorsement Date:** Aug 15, 2011

Maintenance of Endorsement -- Preliminary Analysis

To maintain NQF endorsement endorsed measures are evaluated periodically to ensure that the measures still meets the NQF endorsement criteria ("maintenance"). The emphasis for maintaining endorsement is focused on how effective the measure is for promoting improvements in quality. Endorsed measures should have some experience from the field to inform the evaluation. The emphasis for maintaining endorsement is noted for each criterion.

Criteria 1: Importance to Measure and Report

1a. Evidence
Maintenance measures – less emphasis on evidence unless there is new information or change in evidence since the prior evaluation.

1a. Evidence. The evidence requirements for a *process or intermediate outcome* measure is that it is based on a systematic review (SR) and grading of the body of empirical evidence where the specific focus of the evidence matches

what is being measured.

The developer provides the following evidence for this measure:

- **Systematic Review of the evidence specific to this measure?** ☐ **Yes** ☒ **No**
- **Quality, Quantity and Consistency of evidence provided?** ☐ **Yes** ☒ **No**
- **Evidence graded?** ☐ **Yes** ☒ **No**

Evidence Summary or **Summary of prior review in [year]**

The prior Committee noted that this is measure only assesses visits but not the content of those visits. The Committee agreed that ACOG guidelines recommend a schedule of prenatal visits that are based primarily on expert consensus. The Committee questions the relationship of the visit groups defined in this measure to patient outcomes. The Committee acknowledged that data does show that patients who have no prenatal care have worse outcomes.

Changes to evidence from last review

☒ **The developer attests that there have been no changes in the evidence since the measure was last evaluated.**

☐ **The developer provided updated evidence for this measure:**

Exception to evidence

Since the evidence for this measure is expert consensus rather than empirical evidence, it is insufficient to meet NQF's criterion for evidence. However, an exception to the evidence criterion is allowed if the Committee agrees that empirical evidence is not needed to hold providers accountable for the measure.

Guidance from the Evidence Algorithm

Process measure (Box 1) --> not based on a SR (Box 3) --> no empirical evidence (Box 7) --> systematic assessment of expert opinion (Box 11) --> if Committee agrees it is OK/beneficial to hold providers accountable for performance in the absence of empirical evidence of benefits to patients → rate as INSUFFICIENT WITH EXCEPTION

Staff NOTE: The Committee will first vote on the criterion. If a > 60% votes for insufficient then a second vote will be taken to determine whether the Committee wishes to pass the evidence with exception.

Questions for the Committee:

- *For possible exception to the evidence criterion:*
  - *Are there, or could there be, performance measures of a related health outcome, OR evidence-based intermediate clinical outcomes, intervention/treatment?*
  - *Is there evidence of a systematic assessment of expert opinion beyond those involved in developing the measure?*
  - *Does the SC agree that it is acceptable (or beneficial) to hold providers accountable without empirical evidence?*

Preliminary rating for evidence: ☐ **High** ☐ **Moderate** ☐ **Low** ☒ **Insufficient**

1b. Gap in Care/Opportunity for Improvement and 1b. Disparities
Maintenance measures – increased emphasis on gap and variation

1b. Performance Gap. The performance gap requirements include demonstrating quality problems and opportunity for improvement.

The developer presents the following data for 2013 – 2015:

Frequency of Prenatal Care	2015 (SD) 10th – 90th %ile	2014 (SD) 10th – 90th %ile	2013 (SD) 10th – 90th %ile	2006/2007 data from prior submission
# eligible health plans Median denominator/plan	177 1204	155 1560	137 1148	
less than 21 percent	14% (17%) 2.7% -32%	14% (17%) 1.9% -37%	12% (15%) 2.3%-27%	13.52% / 12.36%
21-40 percent	7.6% (5.7%) 2.5 – 14%	7.6% (56.8%) 2.1 – 17%	5.9% (5.8%) 1.6 – 12%	6.04% / 6.63%
41-60 percent	8.4% (3.6%) 24.4– 13%	8.5% (4.1%) 24.0– 14%	7.7% (4.0%) 3.9 – 13%	7.84% / 7.74%
61-80 percent	14% (5.4%) 7.7 – 21%	14% (5.5%) 5.8 – 21%	14% (5.3%) 7.6 – 21%	14.1% / 13,85%
81+ percent	55% (20%) 27 – 75%	56% (22%) 22 – 78%	61% (19%) 36 – 80%	58.6% / 59.59%

Disparities
Even though the developer presents evidence from the literature that women ages 19 and low-income women are least likely to initiate perinatal care the developer does not collect performance data stratified by age and socioeconomic status.

Questions for the Committee:
- *This measure has been endorsed by NQF for 5 years. How has performance changed over time?*
- *Is there a gap in care that warrants a national performance measure?*
- *Can this measure be used to address disparities?*

Preliminary rating for opportunity for improvement: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments
Criteria 1: Importance to Measure and Report (including 1a, 1b, 1c)

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability

2a1. Reliability Specifications
Maintenance measures – no change in emphasis – specifications should be evaluated the same as with new measures

2a1. Specifications requires the measure, as specified, to produce consistent (reliable) and credible (valid) results about the quality of care when implemented.

Data source(s): Administrative claims, Electronic Clinical Data, Paper Medical Records

Specifications:
- This measure description/specifications indicates that this measure is for Medicaid plans only.
- Two sets of specifications are presented – one for administrative claims and one for medical records.
- The numerator calculation is explained.
  - For the numerator the developer explains "ACOG recommends that women with an uncomplicated

pregnancy receive visits every 4 weeks for the first 28 weeks of pregnancy, every 2–3 weeks until 36 weeks of pregnancy, and weekly thereafter. For example, ACOG recommends 14 visits for a 40-week pregnancy. If the member enrolled during her fourth month (3 missed visits prior to enrollment in the organization), the expected number of visits is 14 – 3 = 11."

- The denominator is "deliveries of live births" in the previous year. Non-live births are excluded.
- The spreadsheet with the value sets include codes for both ICD-9 and ICD-10.
- The developer does not describe any changes to the specifications but indicates that there have been changes.
- A calculation algorithm is included.

Questions for the Committee :

- *Is a Medicaid only measure appropriate?*
- *Are all the data elements clearly defined? Are all appropriate codes included?*
- *Is it appropriate to exclude "non-live births"?*
- *Is the logic or calculation algorithm clear?*
- *Is it likely this measure can be consistently implemented?*

2a2. Reliability Testing Testing attachment
Maintenance measures – less emphasis if no new testing data provided

2a2. Reliability testing demonstrates if the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise enough to distinguish differences in performance across providers.

For maintenance measures, summarize the reliability testing from the prior review:
Measure score reliability was demonstrated through a signal to noise analysis (beta binomial method) for 2010 data:

Frequency of Ongoing Prenatal Care - <21 Percent Rate	0.9897
Frequency of Ongoing Prenatal Care - 21-40 Percent Rate	0.9748
Frequency of Ongoing Prenatal Care - 41-60 Percent Rate	0.9506
Frequency of Ongoing Prenatal Care - 61-80 Percent Rate	0.9533
Frequency of Ongoing Prenatal Care - 81+ Percent Rate	0.9933

This is a commonly used method to assess reliability. The higher the reliability score, the greater is the confidence with which one can distinguish the performance of one plan from another. A reliability score greater than or equal to 0.7 is considered very good.

Describe any updates to testing - none

SUMMARY OF TESTING
Reliability testing level ☒ **Measure score** ☐ **Data element** ☐ **Both**
Reliability testing performed with the data source and level of analysis indicated for this measure ☐ **Yes** ☐ **No**

Method(s) of reliability testing see above

Results of reliability testing see above

Guidance from the Reliability Algorithm
Precise specifications (Box 1) [may be an issue] →empirical reliability testing (Box 2) → measure score testing (Box 4) →appropriate method (Box 5) → High certainty or confidence that the performance measure scores are reliable (Box 6a)

Questions for the Committee:

- *The developer attests there is no new testing for reliability since the last NQF endorsement review. Does the*

Committee agree there is no need for repeat discussion and voting on reliability?

Preliminary rating for reliability: ☒ **High** ☐ **Moderate** ☐ **Low** ☐ **Insufficient**
IF no concerns about two sets of specifications

2b. Validity
Maintenance measures – less emphasis if no new testing data provided

2b1. Validity: Specifications

2b1. Validity Specifications. This section should determine if the measure specifications are consistent with the evidence.

Specifications consistent with evidence in 1a. ☐ **Yes** ☒ **Somewhat** ☐ **No**

Specification not completely consistent with evidence

ACOG guidelines recommend a schedule of prenatal visits, however, the five categories in this measure are not mentioned in the guidelines.

Question for the Committee:

- *Are the specifications consistent with the evidence?*

2b2. Validity testing

2b2. Validity Testing should demonstrate the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality.

For maintenance measures, summarize the validity testing from the prior review: Face validity assessment only by a panel of stakeholders and experts

Describe any updates to validity testing none

SUMMARY OF TESTING

Validity testing level ☒ **Measure score** ☐ **Data element testing against a gold standard** ☐ **Both**

Method of validity testing of the measure score:

- ☒ **Face validity only**
- ☐ **Empirical validity testing of the measure score**

Validity testing method: Face validity only

Validity testing results: "This measure was deemed valid by the expert panel."

The prior Committee asked why there were five categories (number of visits) rather than a simple yes/no criterion where everyone had to meet the same threshold. The Committee was concerned about the variability in reimbursement as a determinant of visit frequency and how case mix in a particular practice influences how a provider would score on this measure and also questioned feasibility of data collection since bundled or global payments are changing billing practices.

Questions for the Committee:

- *Is the test sample adequate to generalize for widespread implementation?*
- *Do the results demonstrate sufficient validity so that conclusions about quality can be made?*
- *Are the five groups included in this measure an indicator of quality?*
- *Do you agree that the score from this measure as specified is an indicator of quality?*

2b3-2b7. Threats to Validity

2b3. Exclusions: No information is provided.

- The only exclusion is non-live births. No data is provided on the frequency of this exclusion.

Questions for the Committee:

- *Are the exclusions consistent with the evidence?*
- *Should "non-live births" be excluded?*
- *Are any patients or patient groups inappropriately excluded from the measure?*
- *Are the exclusions/exceptions of sufficient frequency and variation across providers to be needed (and outweigh the data collection burden)?*

2b4. Risk adjustment: **Risk-adjustment method** ☒ **None** ☐ **Statistical model** ☐ **Stratification**

2b5. Meaningful difference *(can statistically significant and clinically/practically meaningful differences in performance measure scores can be identified)*:

- The developer reports that meaningful differences are determined by "Comparison of means and percentiles; analysis of variance against established benchmarks; if sample size is >400, we would use an analysis of variance."
- See data table in 1b.

Question for the Committee:

- *Does this measure identify meaningful differences about quality?*
- *How is the data from five categories used to compare health plans?*

2b6. Comparability of data sources/methods:

- Two sets of specifications are presented. No information on comparability is discussed.
- No information on the frequency of use of each data source is provided.

Question for the Committee:

- *Are the two sets of specifications likely to produce comparable results?*

2b7. Missing Data The developer reports "NA" for this question.

Guidance from the algorithm:

Consistent with the evidence (Box 1) → Potential threats to validity addressed (Box 2)[comparability of data sources may be an issue] →empirical validity testing (Box 4) → face validity systematically assessed (Box 4) → Highest rating possible is MODERATE if threats to validity adequately addressed

Question for the Committee:

- *The developer attests no new testing for validity has been done since the last NQF endorsement review. Does the Committee agree there is no need for repeat discussion and voting on Evidence?*

Preliminary rating for validity: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**
IF no issues about comparability. In the absence of empirical testing for validity, the highest rating possible is moderate.

Committee pre-evaluation comments

Criteria 2: Scientific Acceptability of Measure Properties (including all 2a, 2b, and 2d)

Criterion 3. Feasibility
Maintenance measures – no change in emphasis – implementation issues may be more prominent

3. Feasibility is the extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

- The developer states that "To allow for widespread reporting across health plans and health care practices, this measure is collected through multiple data sources (administrative data, electronic clinical data, paper records)."
- Measures collected vis administrative claims are quite feasible. The burden of paper medical record review is considerable.

Questions for the Committee:

- *This measure is specified for health plans and integrated delivery systems. Are there any of these entities that would not be collecting this data through claims?*
- *Does the Committee have any concerns about the feasibility of this measure?*

Preliminary rating for feasibility: ☒ **High** ☐ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments
Criteria 3: Feasibility

Criterion 4: Usability and Use
Maintenance measures – increased emphasis – much greater focus on measure use and usefulness, including both impact /improvement and unintended consequences

4. Usability and Use evaluate the extent to which audiences (e.g., consumers, purchasers, providers, policymakers) use or could use performance results for both accountability and performance improvement activities.

Current uses of the measure
Publicly reported? ☒ **Yes** ☐ **No**

Current use in an accountability program? ☒ **Yes** ☐ **No**
OR
Planned use in an accountability program? ☐ **Yes** ☐ **No**

Accountability program details
The developer reports on multiple uses of the measures including public reporting and quality improvement (i.e., NCQA Quality Compass, NCQA State of Health Care Quality Report, Medicaid Child Core Set). The developer provided the following information for each program.

- QUALITY COMPASS: http://www.ncqa.org/tabid/177/Default.aspx This measure is used in Quality Compass which is an indispensable tool used for selecting a health plan, conducting competitor analysis, examining quality improvement and benchmarking plan performance. Provided in this tool is the ability to generate custom reports by selecting plans, measures, and benchmarks (averages and percentiles) for up to three trended years. Results in table and graph formats offer simple comparison of plans' performance against competitors or benchmarks.
- STATE OF HEALTH CARE QUALITY REPORT: This measure is publically reported nationally and by geographic regions in the NCQA State of Health Care annual report. This annual report published by NCQA summarizes findings on quality of care. In 2012 the report included measures on 11.5 million Medicare Advantage beneficiaries in 455 Medicare Advantage health plans, 99.4 million members in 404 commercial health plans, and 14.3 million Medicaid beneficiaries in 136 plans across 50 states.
- MEDICAID CHILD CORE SET: http://www.ncqa.org/tabid/836/Default.aspx These are a core set of health quality

measures for Medicaid-enrolled children. The Medicaid Child Core Set was identified by the Centers for Medicare & Medicaid (CMS) in partnership with the Agency for HealthCare Research and Quality (AHRQ). The data collected from these measures will help CMS to better understand the quality of health care that children enrolled in Medicaid receive nationally. On December 29, 2009, the Secretary posted for public comment in the Federal Register, an initial core set of 24 children´s health care quality measures for voluntary use by Medicaid and CHIP programs. The CHIPRA legislation provides that the Secretary shall issue updates to the Child Core Set beginning in January 2013 and annually thereafter. CMS worked with the National Quality Forum's (NQF) Measures Application Partnership (MAP) to review the Child Core Set and to identify ways to improve it. State data derived from the core measures are part of the Secretary´s Annual Report on the Quality of Care for Children in Medicaid and CHIP.

Improvement results Performance results (data table 1b) show that rates have been steady over the past three years among Medicaid plans.

Unexpected findings (positive or negative) during implementation
The developer reports that "there were no unexpected findings or unexpected benefits."

Potential harms None have been identified.

Feedback :

- In 2015 MAP Medicaid Task Force supported the continued use of this measure in the Medicaid Child Core Set.

***Questions for the Committee**:*

- *Committee members should share any experience with use of this measure in their workplace.*
- *How do Medicaid health plans and systems use this measure?*
- *How useful are the five categories in this measure? How easy is it to track improvement or make comparisons with this measure?*
- *This measure has been in use for many years – what has been the impact?*
- *How do you interpret performance results in which overall performance seems low, yet performance rates in the most recent 3 years are unchanged ?*
- *Is the Committee aware of any unintended consequences or potential harms from the measure? Any unexpected benefits from use of the measure?*
- *Do the benefits of the measure outweigh any potential unintended consequences?*

Preliminary rating for usability and use: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 4: Usability and Use

Criterion 5: Related and Competing Measures

Related or competing measures

Related measure: 1517 : Prenatal & Postpartum Care (PPC)

Harmonization - both measures are from NCQA and are harmonized for use together.

Pre-meeting public and member comments

-

NATIONAL QUALITY FORUM

Measure missing data in MSF 6.5 from MSF 5.0

NQF #: 1391	**NQF Project:** Child Health Quality Measures 2010

1. IMPACT, OPPORTUITY, EVIDENCE - IMPORTANCE TO MEASURE AND REPORT

Importance to Measure and Report is a threshold criterion that must be met in order to recommend a measure for endorsement. All three subcriteria must be met to pass this criterion. See guidance on evidence.

Measures must be judged to be important to measure and report in order to be evaluated against the remaining criteria. **(evaluation criteria)**

1c.1 Structure-Process-Outcome Relationship *(Briefly state the measure focus, e.g., health outcome, intermediate clinical outcome, process, structure; then identify the appropriate links, e.g., structure-process-health outcome; process-health outcome; intermediate clinical outcome-health outcome)***:**

Proper perinatal care is associated with improved birth outcomes. For example, one study found that 25.6 percent of women who did not receive prenatal care delivered preterm infants compared to 9.2 percent of women who received even a minimum amount of prenatal care (Vintzileos et al., 2002).

In 2001, infants of mothers who received no prenatal care had an infant mortality rate of 34.8 per 1,000 live births, compared to an infant mortality rate of only 6.2 per 1,000 when prenatal care was initiated in the first trimester of pregnancy (Matthews et al., 2003). Observational studies have consistently shown that groups having more post-delivery visits have lower maternal, fetal and neonatal illness and mortality.

Regarding postpartum visits, not only do many women experience some degree of emotional liability in the postpartum period, which warrants a follow-up visit, but they will also need personalized care during this time to hasten the development of a healthy mother-infant relationship and a sense of maternal confidence (ACOG, 2002). Should the pregnancy have an abnormal outcome, the postpartum visit is an advantageous time to discuss implications of such conditions as diabetes mellitus, intrauterine growth restriction, preterm birth, hypertension or other conditions that may recur in any future pregnancies (ACOG, 2002). The postpartum visit is also an ideal time to begin preconceptional counseling for patients who may wish to have future pregnancies.

1c.2-3 Type of Evidence *(Check all that apply)***:**

1c.4 Directness of Evidence to the Specified Measure *(State the central topic, population, and outcomes addressed in the body of evidence and identify any differences from the measure focus and measure target population)***:**

The goal of the prenatal contact is to exchange information and identify existing risk factors that may impact the pregnancy

(DoD/VA, 2002). Lack of prenatal care can be considered a high-risk factor for postneonatal death. A study that sought to determine the association between prenatal care (defined as one visit) and postneonatal death rates (defined as the number of deaths of infants between 28 and 365 days of life) found that the postneonatal deaths among women who had prenatal care was 2.1 per 1,000 women, whereas the rate among women without prenatal care was 5.9 per 1,000 (Vintzileos, A, et al., 2002). These rates applied to women without high-risk conditions. Women whose prenatal care fails to meet established standards are at a greater risk for pregnancy complications and negative birth outcomes (National Center for Health Statistics, 1997).

The goal of postpartum care is to assess the physical and psychosocial status of the mother after the mother's discharge. The majority of maternal and neonatal deaths, as well as a significant burden of long-term morbidity, occur during the postpartum period (WHO, 1998). The postpartum visit should include obtaining an interval history and performing a physical exam to evaluate the patient's current status. Additionally, the emotional status of a woman whose pregnancy had an abnormal outcome should be reviewed, as many women experience some degree of emotional liability during the postpartum period. It is also an advisable time to begin preconceptional counseling for patients who may wish to have future pregnancies (ACOG Guidelines, 2002).

1c.5 **Quantity of Studies in the Body of Evidence** *(Total number of studies, not articles)*:

1c.6 **Quality of Body of Evidence** (*Summarize the certainty or confidence in the estimates of benefits and harms to patients across studies in the body of evidence resulting from study factors. Please address: a) study design/flaws; b) directness/indirectness of the evidence to this measure (e.g., interventions, comparisons, outcomes assessed, population included in the evidence); and c) imprecision/wide confidence intervals due to few patients or events):*

1c.7 **Consistency of Results across Studies** *(Summarize the consistency of the magnitude and direction of the effect):*

1c.8 **Net Benefit** *(Provide estimates of effect for benefit/outcome; identify harms addressed and estimates of effect; and net benefit - benefit over harms)*:

1c.9 **Grading of Strength/Quality of the Body of Evidence**. Has the body of evidence been graded?

1c.10 **If body of evidence graded, identify the entity that graded the evidence including balance of representation and any disclosures regarding bias:**

1c.11 **System Used for Grading the Body of Evidence:** Expert Consensus

1c.12 **If other, identify and describe the grading scale with definitions:**

1c.13 **Grade Assigned to the Body of Evidence:** Good

1c.14 Summary of Controversy/Contradictory Evidence: None

1c.15 Citations for Evidence other than Guidelines*(Guidelines addressed below)*:

Veterans Health Administration, Department of Defense. DoD/VA clinical practice guideline for the management of uncomplicated pregnancy. Washington (DC): Department of Veterans Affairs; 2002 October.

Department of Reproductive Health and Research (RHR), World Health Organization (WHO). Postpartum care of the mother and newborn: a practical guide; 1998.

Vintzileos, A, Ananth, C, Smulian, JC, Scorza, WE, Knuppel, RA. The impact of prenatal care on postneonatal deaths in the presence and absence of antenatal high-risk conditions. Am J Obstet Gynecol November 2002; 187(5):1258-1262

1c.16 Quote verbatim, <u>the specific guideline recommendation</u> *(Including guideline # and/or page #)*:

ACOG

Timeliness of Prenatal Care: Once pregnancy occurs, patients should have early contact with an obstetrician to begin counseling about prenatal testing and to develop a management plan. The frequency of subsequent antepartum office visits is determined by the individual needs of the woman and the assessment of her risks. According to ACOG guidelines, for an uncomplicated pregnancy, the ACOG guidelines suggest the following frequency of office visits: monthly office visits from the initial prenatal visit until 29 weeks of pregnancy; weekly office visits from 36 weeks until delivery; office visits every two to three weeks from 29 weeks to 36 weeks of pregnancy. (ACOG guidelines, 2010)

Postpartum Care: The timing of the postpartum visit has been a topic for debate. According to the ACOG, the mother should visit her physician for a postpartum review and examination approximately 4 to 6 weeks after delivery. This interval may be modified according to the needs of the patient with medical, obstetric, or intercurrent complications (ACOG Guidelines, 2002).

The DoD/VA clinical practice guideline for management of uncomplicated pregnancy

Spports the recommended 8 weeks after delivery postpartum visit. Evidence suggests that eight weeks is the optimal time to decrease the rate of false positive cervical smears, though consideration of the mother's schedule should also be taken into account (2002).

A visit within 7-14 days of delivery may be advisable after a cesarean delivery or a complicated gestation, primarily to assess the surgical wounds and healing. The standard postpartum care visit is recommended in follow-up to this initial visit (ACOG Guidelines) to ensure the woman's uterus has reduced to its normal size, and to conduct a depression screening and family planning counseling.

1c.17 Clinical Practice Guideline Citation: American Academy of Pediatrics and The American College of Obstetricians and Gynecologists. Guidelines for Perinatal Care (5th Edition). October 2002.

1c.18 National Guideline Clearinghouse or other URL: Routine prenatal and postnatal care. http://www.guideline.gov/content.aspx?id=13174&search=prenatal+and+postpartum+care

1c.19 Grading of Strength of Guideline Recommendation. Has the recommendation been graded?

1c.20 If guideline recommendation graded, identify the entity that graded the evidence including balance of representation and any disclosures regarding bias:

1c.21 System Used for Grading the Strength of Guideline Recommendation: USPSTF

1c.22 If other, identify and describe the grading scale with definitions:

1c.23 Grade Assigned to the Recommendation: Good

1c.24 Rationale for Using this Guideline Over Others: The measures are access and use of service measures that are based on the body of evidence and guidelines regarding perinatal care.

Based on the NQF descriptions for rating the evidence, what was the <u>developer's assessment</u> of the quantity, quality, and consistency of the body of evidence?

1c.25 Quantity: **1c.26** Quality: **1c.27** Consistency**:**

1. Evidence, Performance Gap, Priority – Importance to Measure and Report

Extent to which the specific measure focus is evidence-based, important to making significant gains in healthcare quality, and improving health outcomes for a specific high-priority (high-impact) aspect of healthcare where there is variation in or overall less-than-optimal performance. ***Measures must be judged to meet all subcriteria to pass this criterion and be evaluated against the remaining criteria.***

1a. Evidence to Support the Measure Focus – See attached Evidence Submission Form

1391_Evidence_MSF5.0_Data-635278481500031458-635787040722965660.doc

1b. Performance Gap

Demonstration of quality problems and opportunity for improvement, i.e., data demonstrating:

- considerable variation, or overall less-than-optimal performance, in the quality of care across providers; and/or
- disparities in care across population groups.

1b.1. **Briefly explain the rationale for this measure** *(e.g., the benefits or improvements in quality envisioned by use of this measure)*
This measure assesses appropriate prenatal care by seeking to ensure that pregnant women receive the proper number of prenatal care visits for each stage of a pregnancy. Women who utilize prenatal care can minimize their risk for pregnancy complications and negative birth outcomes, which include higher risk of infant morbidity and mortality. Adherence to this measure could improve access to services and provide opportunities to improve health of both the infant and the mother.

1b.2. **Provide performance scores on the measure as specified (current and over time) at the specified level of analysis.** *(This is required for endorsement maintenance. Include mean, std dev, min, max, interquartile range, scores by decile. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included). This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*
The following data are extracted from HEDIS data collection reflecting the most recent years of measurement for this measure. Performance data are at the health plan level and summarized by mean, standard deviation, minimum, maximum, and performance at the 10th, 25th, 50th, 75th and 90th percentile. Data are stratified by year.

Frequency of Prenatal Care - less than 21 percent

YEAR| MEAN | ST DEV | 10TH | 25TH | 50TH | 75TH | 90TH | Interquartile Range

2015 | 14% | 17% | 2.7% | 5.6% | 8.6% | 15% | 32% | 9.5%

2014 | 14% | 17% | 1.9% | 4.9% | 9.3% | 16% | 37% | 11%

2013 | 12% | 15% | 2.3% | 4.2% | 8.0% | 14% | 27% | 9.5%

Frequency of Prenatal Care – 21-40 percent

YEAR| MEAN | ST DEV | 10TH | 25TH | 50TH | 75TH | 90TH | Interquartile Range

2015 | 7.6% | 5.7% | 2.5% | 3.8% | 5.6% | 10% | 14% | 6.7%

2014 | 7.6% | 6.8% | 2.1% | 3.5% | 5.4% | 8.8% | 17% | 5.4%

2013 | 5.9% | 5.8% | 1.6% | 2.8% | 4.2% | 6.6% | 12% | 3.8%

Frequency of Prenatal Care – 41-60 percent

YEAR| MEAN | ST DEV | 10TH | 25TH | 50TH | 75TH | 90TH | Interquartile Range

2015 | 8.4% | 3.6% | 4.4% | 5.9% | 7.8% | 10% | 13% | 4.3%

2014 | 8.5% | 4.1% | 4.0% | 5.6% | 7.7% | 11% | 14% | 5.2%

2013 | 7.7% | 4.0% | 3.9% | 4.9% | 6.8% | 9.5% | 13% | 4.6%

Frequency of Prenatal Care – 61-80 percent

YEAR| MEAN | ST DEV | 10TH | 25TH | 50TH | 75TH | 90TH | Interquartile Range

2015 | 14% | 5.4% | 7.7% | 12% | 14% | 18% | 21% | 5.9%

2014 | 14% | 5.5% | 5.8% | 11% | 15% | 17% | 21% | 6.3%

2013 | 14% | 5.3% | 7.6% | 11% | 13% | 16% | 21% | 5.6%

Frequency of Prenatal Care – 81+ percent

YEAR| MEAN | ST DEV | 10TH | 25TH | 50TH | 75TH | 90TH | Interquartile Range

2015 | 55% | 20% | 27% | 47% | 59% | 70% | 75% | 23%

2014 | 56% | 22% | 22% | 44% | 60% | 71% | 78% | 28%

2013 | 61% | 19% | 36% | 51% | 65% | 74% | 80% | 23%

Data for this measure are from the Healthcare Effectiveness Data and Information Set (HEDIS). In 2013, HEDIS measures covered more than 171 million people from 814 HMOs and 353 PPOs. Below is a description of the denominator for this measure.

Below are the number of health plans and the median eligible population for this measure for the three years of performance data presented.

Frequency of Prenatal Care - less than 21 percent

YEAR | N Plans | Median Denominator Size per plan

2015 | 177 | 1204

2014 | 155 | 1560

2013 | 137 | 1148

Frequency of Prenatal Care – 21-40 percent

YEAR | N Plans | Median Denominator Size per plan

2015 | 177 | 1204

2014 | 155 | 1560

2013 | 137 | 1148

Frequency of Prenatal Care – 41-60 percent

YEAR | N Plans | Median Denominator Size per plan

2015 | 177 | 1204

2014 | 155 | 1560

2013 | 137 | 1148

Frequency of Prenatal Care – 61-80 percent

YEAR | N Plans | Median Denominator Size per plan

2015 | 177 | 1204

2014 | 155 | 1560

2013 | 137 | 1148

Frequency of Prenatal Care – 81+ percent

YEAR | N Plans | Median Denominator Size per plan

2015 | 177 | 1204

2014 | 155 | 1560

2013 | 137 | 1148

1b.3. If no or limited performance data on the measure as specified is reported in 1b2, then provide a summary of data from the literature that indicates opportunity for improvement or overall less than optimal performance on the specific focus of measurement.
NA

1b.4. Provide disparities data from the measure as specified (current and over time) by population group, e.g., by race/ethnicity, gender, age, insurance status, socioeconomic status, and/or disability. *(This is required for endorsement maintenance. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities include.) This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*
This measure is not stratified by variables to detect disparities. NCQA does not currently collect performance data stratified by race, ethnicity, or language. Escarce et al. have described in detail the difficulty of collecting valid data on race, ethnicity and language at the health plan level (Escarce, 2011). While not specified in the measure, this measure can also be stratified by demographic variables, such as race/ethnicity or socioeconomic status, in order to assess the presence of health care disparities. The HEDIS Health Plan Measure Set contains two measures that can assist with stratification to assess health care disparities. The Race/Ethnicity Diversity of Membership and the Language Diversity of Membership measures were designed to promote standardized methods for collecting these data and follow Office of Management and Budget and Institute of Medicine guidelines for collecting and categorizing race/ethnicity and language data. In addition, NCQA's Multicultural Health Care Distinction Program outlines standards for collecting, storing and using race/ethnicity and language data to assess health care disparities. Based on extensive work by NCQA to understand how to promote culturally and linguistically appropriate services among plans and providers, we have many examples of how health plans have used HEDIS measures to design quality improvement programs to decrease disparities in care.

1b.5. If no or limited data on disparities from the measure as specified is reported in 1b4, then provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement. Include citations.
Researchers have explored disparities in prenatal care and low birth rates, a potential consequence of lack of timely perinatal care. Low-income women have been found to enroll in health plans and initiate prenatal care at later stages of pregnancy than women of higher incomes. In addition, women ages 19 and under are least likely to initiate prenatal care as compared to older women. Delayed enrollment and care initiation results in limited access to prenatal screenings and other resources available within the Medicaid program for the duration of the pregnancy. One study concluded that early establishment of prenatal care is important to identifying potential birth risks and implementing an appropriate care plan during pregnancy (Egerter S et al 2002).

Although birth outcomes have been improving as a whole for the U.S. population, these trends are not as pronounced among certain racial minorities. In comparison to those for White and Hispanic populations, rates of preterm birth and low birth weight are significantly higher for the Black population. Preterm birth rate for the Black population was 16.8 percent, compared to 10.5 percent for non-Hispanic Whites and 11.7 percent for the Hispanic population (CDC 2011). This disparity is more evident for early preterm births (infants born before 32 weeks gestation): the Black population has twice as many early preterm births compared to their non-Hispanic White and Hispanic counterparts (CDC 2013). Complications and poorer health outcomes for pre-term birth infants extend beyond increased hospital stays and care needs at birth. Preterm birth infants are at an increased risk of suffering from more long-term conditions such as asthma or experiencing digestion difficulties in early life (IOM 2007).

1c. High Priority (previously referred to as High Impact)

The measure addresses:

- a specific national health goal/priority identified by DHHS or the National Priorities Partnership convened by NQF; OR
- a demonstrated high-priority (high-impact) aspect of healthcare (e.g., affects large numbers of patients and/or has a substantial impact for a smaller population; leading cause of morbidity/mortality; high resource use (current and/or future); severity of illness; and severity of patient/societal consequences of poor quality).

1c.1. **Demonstrated high priority aspect of healthcare**

Affects large numbers, High resource use, Patient/societal consequences of poor quality

1c.2. **If Other:**

1c.3. **Provide epidemiologic or resource use data that demonstrates the measure addresses a high priority aspect of healthcare. List citations in 1c.4.**

Research indicates that early, comprehensive prenatal care can promote healthier pregnancies and reduce the risk of some costly, adverse birth outcomes. Prenatal care allows providers to address key risk factors associated with low birth as smoking, alcohol use, proper weight gain and nutrition and management of existing chronic diseases (CDC 2012). Lack of prenatal care can lead to complications such as preterm birth, which has both negative health consequences and financial impact on the health care system and patients. One study found that 25.6 percent of women who did not receive prenatal care delivered preterm infants compared to nine percent of women who received even a minimum amount of prenatal care (Vintzileos A et al. 2002).

Data from the Centers for Disease Control and Prevention (CDC) indicate 1 in 8 infants are born preterm each year in the United States. These preterm births represented 35 percent of total infant deaths during 2009 (CDC 2013). Children born prematurely are more likely to suffer a variety of motor, cognitive, academic, and behavioral problems, as well as physical health problems in adulthood, such as increased rates of coronary heart disease, stroke, hypertension, and non-insulin dependent diabetes (PHC4 2003; CDC 2013). Preterm infants can experience feeding difficulties resulting from under-developed gastrointestinal systems. Low birth weight babies may face serious health problems upon birth but also are at increased risk of long-term disabilities (March of Dimes 2013).

Premature babies are a driver of excessive maternity costs. Research indicates that early, comprehensive prenatal care can promote healthier pregnancies and reduce the risk of costly, adverse birth outcomes (National Business Group on Health, 2009). A 2007 report by the Institute of Medicine (IOM) indicates preterm birth costs the U.S. $26.2 billion each year. These costs are primarily related to care for the infant but also encompass $1.9 billion in care provided during labor and delivery for the mother (March of Dimes 2013). Moreover, pregnancy complications before delivery account for more than two million hospital days of care each year (CDC 2008).

High costs associated with poor birth outcomes extend beyond the early years of life and into adulthood. Expenditures related to preterm birth have been connected with programs and education services to support preterm infants with developmental delays. A 2007 report from the IOM valued the cost of lost work and pay for persons born prematurely to be $5.6 billion. This amount represents the long-term effects of preterm birth on a person's education and work capabilities (March of Dimes 2013).

The IOM estimates that for every dollar spent on prenatal care, $3.38 is saved in medical costs for low birth weight babies (PHC4 2003). Adequate and timely prenatal care not only improves the likelihood of a healthy birth but also provides a higher probability of a prosperous future for the infant. The societal costs of prematurity add greater incentive to ensure women are initiating and receiving adequate and timely prenatal care.

1c.4. **Citations for data demonstrating high priority provided in 1a.3**

1c.5. **If a PRO-PM (e.g. HRQoL/functional status, symptom/burden, experience with care, health-related behaviors), provide evidence that the target population values the measured PRO and finds it meaningful.** *(Describe how and from whom their input was obtained.)*

2. Reliability and Validity—Scientific Acceptability of Measure Properties

Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. ***Measures must be judged to meet the subcriteria for both reliability and validity to pass this criterion and be evaluated against the remaining criteria.***

2a.1. Specifications The measure is well defined and precisely specified so it can be implemented consistently within and across organizations and allows for comparability. eMeasures should be specified in the Health Quality Measures Format (HQMF) and the Quality Data Model (QDM).

De.5. **Subject/Topic Area** *(check all the areas that apply):*
Perinatal and Reproductive Health, Perinatal and Reproductive Health : Perinatal, Perinatal and Reproductive Health : Screening

De.6. **Cross Cutting Areas** *(check all the areas that apply):*
Access

S.1. **Measure-specific Web Page** *(Provide a URL link to a web page specific for this measure that contains current detailed specifications including code lists, risk model details, and supplemental materials. Do not enter a URL linking to a home page or to general information.)*
NA

S.2a. **If this is an eMeasure**, HQMF specifications must be attached. Attach the zipped output from the eMeasure authoring tool (MAT) - if the MAT was not used, contact staff. (Use the specification fields in this online form for the plain-language description of the specifications)
This is not an eMeasure **Attachment:**

S.2b. **Data Dictionary, Code Table, or Value Sets** *(and risk model codes and coefficients when applicable) must be attached. (Excel or csv file in the suggested format preferred - if not, contact staff)*
Attachment **Attachment:** 1391_FPC_Value_Sets.xlsx

S.3. **For endorsement maintenance**, please briefly describe any changes to the measure specifications since last endorsement date and explain the reasons.

S.4. **Numerator Statement** *(Brief, narrative description of the measure focus or what is being measured about the target population, i.e., cases from the target population with the target process, condition, event, or outcome)*
IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
Women who had the appropriate number of expected prenatal visits

S.5. **Time Period for Data** *(What is the time period in which data will be aggregated for the measure, e.g., 12 mo, 3 years, look back to August for flu vaccination? Note if there are different time periods for the numerator and denominator.)*
The measurement year (12 month period).

S.6. **Numerator Details** *(All information required to identify and calculate the cases from the target population with the target process, condition, event, or outcome such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
IF an OUTCOME MEASURE, describe how the observed outcome is identified/counted. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
Administrative Specifications

Women who had an unduplicated count of less than 21 percent, 21 percent–40 percent, 41 percent–60 percent, 61 percent–80 percent or greater than or equal to 81 percent of the number of expected visits, adjusted for the month of pregnancy at time of enrollment and gestational age. For each delivery, follow the steps below to calculate each woman's ratio of observed-to-expected prenatal care visits.

Step 1: Identify the delivery date using hospital discharge data.

Step 2: Identify the date when the member enrolled in the organization and determine the stage of pregnancy at time of enrollment. If the member has gaps in enrollment during pregnancy, use the last enrollment segment to determine continuous enrollment in the organization. For members with a gap in enrollment any time during pregnancy (including a gap in the first trimester), the last enrollment segment is the enrollment start date during the pregnancy that is closest to the delivery date.

Use the following approach (or an equivalent method) to calculate the stage of pregnancy at time of enrollment. If gestational age is not available, assume a gestational age of 280 days (40 weeks).

• Convert gestational age into days.

• Subtract gestational age (in days) from the date of delivery (step 1).

• Subtract the date obtained above from the date when the member enrolled in the organization to determine the stage of pregnancy at time of enrollment.

• Divide the numbers of days the member was pregnant at enrollment (step 3) by 30. Round the resulting number according to the .5 rule to a whole number.

For example, delivery date is August 8, 2015; gestational age is 33 weeks; date of enrollment is May 6, 2015. Given these variables, the process is:

– Gestational age in days is 231 days (33 weeks ´ 7 days/week).

– Date of delivery – gestational age (in days) is December 20, 2014 (August 8, 2015 – 231 days).

– Date when the member enrolled in the organization – date obtained in step 2 is 137 days (May 6, 2015 – December 20, 2014).

– Month in which prenatal care began is 4.56 months (137 days/30 days) and then round up to 5 months using the 0.5 rule.

This member's stage of pregnancy at time of enrollment is 5 months.

Step 3: Use Table FPC-A to find the number of recommended prenatal visits by gestational age and stage of pregnancy at time of enrollment per the American College of Obstetricians and Gynecologists (ACOG). The chart subtracts the number of missed visits prior to the date the member enrolled from the number of recommended visits for a given gestational age.

ACOG recommends that women with an uncomplicated pregnancy receive visits every

4 weeks for the first 28 weeks of pregnancy, every 2–3 weeks until 36 weeks of pregnancy, and weekly thereafter. For example, ACOG recommends 14 visits for a 40-week pregnancy. If the member enrolled during her fourth month (3 missed visits prior to enrollment in the organization), the expected number of visits is 14 – 3 = 11.

For deliveries with a gestational age less than 28 weeks or >43 weeks, calculate the expected number of prenatal care visits using the date when the member enrolled and ACOG's recommended schedule of visits. For example, if gestational age is 26 weeks and the member enrolled during her second month of pregnancy, the expected number of prenatal care visits is 5 (6 expected visits [1 visit every 4 weeks or 6 visits in 24 weeks], less 1 visit missed in the first month).

If gestational age is 44 weeks and the member enrolled during her third month of pregnancy, the expected number of prenatal care visits is 16 (14 expected visits for a 40-week gestation plus 1 visit each additional week [18 total expected prenatal care visits], less 2 visits missed in the first and second months).

Step 4: Identify the number of discrete prenatal care visits the member received during the course of her pregnancy and while enrolled in the organization using claims and encounter data.

To identify prenatal visits that occurred during the first trimester, refer to the Prenatal and Postpartum Care measure decisions rules for Identifying Prenatal Care For Women Continuously Enrolled During the First Trimester.

To identify prenatal visits that occurred during the second and third trimester, refer to the prenatal and postpartum care measure

instructions for Identifying Prenatal Care For Women Not Continuously Enrolled During the First Trimester. Visits that occur on the date of delivery and meet the prenatal visit criteria count toward the measure.

All criteria must be met for encounters to be counted as a discrete prenatal care visit. For example, Decision Rules 2 and 3 require multiple components (typically a visit combined with a diagnosis code or another prenatal service such as a lab test or an ultrasound). Ultrasound and lab results alone are not considered a discrete prenatal care visit unless they are combined with other criteria.

Services that occur over multiple visits can be combined to create a discrete prenatal care visit if all services occur within the time frame established in the measure and services are not double counted. Organizations must develop systems to avoid double counting. For example, a code from the Stand Alone Prenatal Visits Value Set on the same date of service as a code from the Prenatal Visits Value Set is interpreted to represent a single visit/encounter and may not be counted twice. If the member had a gap in enrollment, count only the visits received during the last enrollment segment.

Step 5: Calculate the ratio of observed visits (step 4) to expected visits (step 3).

Step 6: Report each woman in the appropriate category:

• less than 21 percent.

• 21 percent–40 percent.

• 41 percent–60 percent.

• 61 percent–80 percent.

• greater than or equal to 81 percent of expected visits.

Medical Record Specification

Women who had an unduplicated count of the number of expected visits that was less than 21 percent, 21 percent–40 percent, 41 percent–60 percent, 61 percent–80 percent or greater than or equal to 81 percent of the number of expected visits, adjusted for the month of pregnancy at time of enrollment and gestational age. The visits may be identified through either administrative data or medical record review.

The numerator is calculated retroactively from date of delivery or EDD.

Use the medical record documentation requirements in the Prenatal and Postpartum Care measure to identify prenatal visits that occur during the first, second and third trimesters.

Identify gestational age at birth from the hospital record (e.g., admission write-ups, histories and physicals, discharge summaries or labor and delivery records) or birth certificate. Gestational age is the number of completed weeks that elapsed between the first day of the last normal menstrual period and the date of delivery. If gestational age is not available, assume a gestational age of 280 days (40 weeks).

Methods recommended to determine gestational age are:

Physician ascertainment using ultrasound or Dubowitz assessment.

Last menstrual period (LMP) calculation (date of LMP – date of delivery) divided by 7. If gestational age is recorded or calculated in fractions of a week, round down to the lower whole number.

S.7. Denominator Statement *(Brief, narrative description of the target population being measured)*

The percentage of deliveries of live births between November 6 of the year prior to the measurement year and November 5 of the measurement year.

S.8. **Target Population Category** *(Check all the populations for which the measure is specified and tested if any):*
Maternal Health

S.9. **Denominator Details** *(All information required to identify and calculate the target population/denominator such as definitions, specific data collection items/responses , code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
Product Line: Medicaid.

Continuous enrollment: 43 days prior to delivery through 56 days after delivery.

Allowable gap: No allowable gap during the continuous enrollment period.

Anchor date: Date of delivery.

Benefit: Medical.

Event/ diagnosis: Delivered a live birth on or between November 6 of the year prior to the measurement year and November 5 of the measurement year. Include women who delivered in any setting.

Multiple births. Women who had two separate deliveries (different dates of service) between November 6 of the year prior to the measurement year and November 5 of the measurement year are counted twice. Women who had multiple live births during one pregnancy are counted once.

Follow the steps below to identify the eligible population, which is the denominator for both rates.

Step 1: Identify deliveries. Identify all women with a delivery (Deliveries Value Set) between November 6 of the year prior to the measurement year and November 5 of the measurement year.

Step 2: Exclude non-live births (Non-live Births Value Set). Step 3: Identify continuous enrollment. Determine if enrollment was continuous between 43 days prior to delivery and 56 days after delivery, with no gaps.

S.10. **Denominator Exclusions** *(Brief narrative description of exclusions from the target population)*
Exclude non-live births

S.11. **Denominator Exclusion Details** *(All information required to identify and calculate exclusions from the denominator such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
See corresponding Excel document for the Non-live Births Value Set

S.12. **Stratification Details/Variables** *(All information required to stratify the measure results including the stratification variables, definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format with at S.2b)*
None

S.13. **Risk Adjustment Type** (Select type. Provide specifications for risk stratification in S.12 and for statistical model in S.14-15)
No risk adjustment or risk stratification
If other:

S.14. **Identify the statistical risk model method and variables** *(Name the statistical method - e.g., logistic regression and list all the risk factor variables. Note - risk model development and testing should be addressed with measure testing under Scientific Acceptability)*
NA

S.15. **Detailed risk model specifications** *(must be in attached data dictionary/code list Excel or csv file. Also indicate if available at measure-specific URL identified in S.1.)*
Note: Risk model details (including coefficients, equations, codes with descriptors, definitions), should be provided on a separate

worksheet in the suggested format in the Excel or csv file with data dictionary/code lists at S.2b.

S.15a. **Detailed risk model specifications** (*if not provided in excel or csv file at S.2b*)
NA

S.16. **Type of score:**
Rate/proportion
If other:

S.17. **Interpretation of Score** *(Classifies interpretation of score according to whether better quality is associated with a higher score, a lower score, a score falling within a defined interval, or a passing score)*
Better quality = Higher score

S.18. **Calculation Algorithm/Measure Logic** *(Describe the calculation of the measure score as an ordered sequence of steps including identifying the target population; exclusions; cases meeting the target process, condition, event, or outcome; aggregating data; risk adjustment; etc.)*
Step 1: Calculate the eligible population following the instructions in the denominator details listed in section S.9.

Step 2: Remove the exclusions identified in section S.10.

Step 3: Calculate the numerator following the instructions in the numerator details listed in section S.6.

Step 4: Divide the numerator from Step 3 by the denominator from Step 2 to determine the rate.

S.19. **Calculation Algorithm/Measure Logic Diagram URL or Attachment** *(You also may provide a diagram of the Calculation Algorithm/Measure Logic described above at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
No diagram provided

S.20. **Sampling** *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*
IF a PRO-PM, identify whether (and how) proxy responses are allowed.
A systematic sample of members drawn from the eligible population. Frequency of Ongoing Prenatal Care and Prenatal and Postpartum Care measures must use the same systematic sample for both. The organization may reduce the sample size using the current year's lowest product-line-specific administrative rate for the rate of women who received >=81 percent of expected prenatal care visits and the two rates from Prenatal and Postpartum Care. It may also use the prior year's lowest audited product-line-specific rates for the rate of women who received >=81 percent of expected prenatal care visits and the two rates from Prenatal and Postpartum Care.

S.21. **Survey/Patient-reported data** *(If measure is based on a survey, provide instructions for conducting the survey and guidance on minimum response rate.)*
IF a PRO-PM, specify calculation of response rates to be reported with performance measure results.
NA

S.22. **Missing data** (specify how missing data are handled, e.g., imputation, delete case.)
Required for Composites and PRO-PMs.
NA

S.23. **Data Source** *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*
If other, please describe in S.24.
Administrative claims, Electronic Clinical Data, Paper Medical Records

S.24. **Data Source or Collection Instrument** *(Identify the specific data source/data collection instrument e.g. name of database, clinical registry, collection instrument, etc.)*
IF a PRO-PM, identify the specific PROM(s); and standard methods, modes, and languages of administration.
This measure is based on administrative claims and medical record documentation collected in the course of providing care to health plan members. NCQA collects the Healthcare Effectiveness Data and Information Set (HEDIS) data for this measure directly

from Health Management Organizations and Preferred Provider Organizations via NCQA's online data submission system.

S.25. **Data Source or Collection Instrument** *(available at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
No data collection instrument provided

S.26. **Level of Analysis** *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*
Health Plan, Integrated Delivery System

S.27. **Care Setting** *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*
Ambulatory Care : Clinician Office/Clinic
If other:

S.28. **COMPOSITE Performance Measure** - Additional Specifications *(Use this section as needed for aggregation and weighting rules, or calculation of individual performance measures if not individually endorsed.)*
NA

2a. Reliability – See attached Measure Testing Submission Form
2b. Validity – See attached Measure Testing Submission Form
1391_MeasureTesting_MSF5.0_Data-635278481500031458-635787040724057660.doc

2. RELIABILITY & VALIDITY - SCIENTIFIC ACCEPTABILITY OF MEASURE PROPERTIES

Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. **(evaluation criteria)**

Measure testing must demonstrate adequate reliability and validity in order to be recommended for endorsement. Testing may be conducted for data elements and/or the computed measure score. Testing information and results should be entered in the appropriate field. Supplemental materials may be referenced or attached in item 2.1. See guidance on measure testing.

2a2. Reliability Testing. (*Reliability testing was conducted with appropriate method, scope, and adequate demonstration of reliability*.)

2a2.1 Data/Sample *(Description of the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)***:**

Medicaid plans who reported the measure in 2010

2a2.2 Analytic Method *(Describe method of reliability testing & rationale)***:**

Reliability was estimated by using the beta-binomial model. Beta-binomial is a better fit when estimating the reliability of simple pass/fail rate measures as is the case with most HEDIS® health plan measures. The beta-binomial model assumes the plan score is a binomial random variable conditional on the plan´s true value that comes from the beta distribution. The beta distribution is usually defined by two parameters, alpha and beta. Alpha and beta can be thought of as intermediate calculations to get to the needed variance estimates. The beta distribution can be symmetric, skewed or even U-shaped.

Reliability used here is the ratio of signal to noise. The signal in this case is the proportion of the variability in measured performance that can be explained by real differences in performance. A reliability of zero implies that all the variability in a measure is attributable to measurement error. A reliability of one implies that all the variability is attributable to real differences in performance. The higher the reliability score, the greater is the confidence with which one can distinguish the performance of one plan from another. A reliability score greater than or equal to 0.7 is considered very good.

2a2.3 Testing Results *(Reliability statistics, assessment of adequacy in the context of norms for the test conducted)***:**

Frequency of Ongoing Prenatal Care - <21 Percent Rate

0.9897

Frequency of Ongoing Prenatal Care - 21-40 Percent Rate

0.9748

Frequency of Ongoing Prenatal Care - 41-60 Percent Rate

0.9506

Frequency of Ongoing Prenatal Care - 61-80 Percent Rate

0.9533

Frequency of Ongoing Prenatal Care - 81+ Percent Rate

0.9933

2b. VALIDITY. Validity, Testing, including all Threats to Validity: H☐ M☐ L☐ I ☐

2b1.1 Describe how the measure specifications *(measure focus, target population, and exclusions)* **are consistent with**

the evidence cited in support of the measure focus (*criterion 1c)* **and identify any differences from the evidence:**

2b2. Validity Testing. (*Validity testing was conducted with appropriate method, scope, and adequate demonstration of validity.)*

2b2.1 Data/Sample *(Description of the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)***:**

expert panel

2b2.2 Analytic Method *(Describe method of validity testing and rationale; if face validity, describe systematic assessment)***:**

NCQA tested the measure for face validity using a panel of stakeholders with specific expertise in measurement and child health care. This panel included representatives from key stakeholder groups, including pediatricians, family physicians, health plans, state Medicaid agencies and researchers. Experts reviewed the results of the field test and assessed whether the results were consistent with expectations, whether the measure represented quality care, and whether we were measuring the most important aspect of care in this area.

2b2.3 Testing Results *(Statistical results, assessment of adequacy in the context of norms for the test conducted; if face validity, describe results of systematic assessment)***:**

This measure was deemed valid by the expert panel.

POTENTIAL THREATS TO VALIDITY. (*All potential threats to validity were appropriately tested with adequate results.*)

2b3. Measure Exclusions. (*Exclusions were supported by the clinical evidence in 1c or appropriately tested with results demonstrating the need to specify them.*)

2b3.1 Data/Sample for analysis of exclusions *(Description of the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)***:**

NA

2b3.2 Analytic Method (*Describe type of analysis and rationale for examining exclusions, including exclusion related to patient preference)***:**

NA

2b3.3 Results *(Provide statistical results for analysis of exclusions, e.g., frequency, variability, sensitivity analyses):*

NA

2b4. Risk Adjustment Strategy. (*For outcome measures, adjustment for differences in case mix (severity) across measured entities was appropriately tested with adequate results.*)

2b4.1 Data/Sample *(Description of the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)***:**

NA

2b4.2 Analytic Method (*Describe methods and rationale for development and testing of risk model or risk stratification including selection of factors/variables)***:**

NA

2b4.3 Testing Results *(Statistical risk model: Provide quantitative assessment of relative contribution of model risk factors; risk model performance metrics including cross-validation discrimination and calibration statistics, calibration curve and risk decile plot, and assessment of adequacy in the context of norms for risk models. Risk stratification: Provide quantitative assessment of relationship of risk factors to the outcome and differences in outcomes among the strata)***:**

NA

2b4.4 If outcome or resource use measure is not risk adjusted, provide rationale and analyses to justify lack of adjustment: The measure assesses utilization and access in a general population; risk adjustment is not indicated.

2b5. Identification of Meaningful Differences in Performance. (*The performance measure scores were appropriately analyzed and discriminated meaningful differences in quality.*)

2b5.1 Data/Sample *(Describe the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)***:**

HEDIS National Data

2b5.2 Analytic Method *(Describe methods and rationale to identify statistically significant and practically/meaningfully differences in performance)***:**

Comparison of means and percentiles; analysis of variance against established benchmarks; if sample size is >400, we would use an analysis of variance

2b5.3 Results *(Provide measure performance results/scores, e.g., distribution by quartile, mean, median, SD, etc.; identification of statistically significant and meaningfully differences in performance)*:

Frequency of Ongoing Prenatal Care

National Means

HEDIS 2006 Data

<21%

13.52

HEDIS 2007 Data

12.36

>= 81%

HEDIS 2006 Data

58.6

HEDIS 2007 Data

59.59

21-40%

HEDIS 2006 Data

6.04

HEDIS 2007 Data

6.63

41-60%

HEDIS 2006 Data

7.84

HEDIS 2007 Data

7.74

61-80%

HEDIS 2006 Data

14.1

HEDIS 2007 Data

13.85

2b6. Comparability of Multiple Data Sources/Methods. (*If specified for more than one data source, the various approaches result in comparable scores.*)

2b6.1 Data/Sample *(Describe the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)***:**

2b6.2 Analytic Method *(Describe methods and rationale for testing comparability of scores produced by the different data sources specified in the measure)***:**

2b6.3 Testing Results *(Provide statistical results, e.g., correlation statistics, comparison of rankings; assessment of adequacy in the context of norms for the test conducted)***:**

2c. Disparities in Care: H☐ M☐ L☐ I ☐ NA☐ (*If applicable, the measure specifications allow identification of disparities.*)

2c.1 If measure is stratified for disparities, provide stratified results *(Scores by stratified categories/cohorts)***:** The measure is not stratified to detect disparities.

2c.2 If disparities have been reported/identified (e.g., in 1b), but measure is not specified to detect disparities, please explain:

NA
2.1-2.3 Supplemental Testing Methodology Information:
Steering Committee: Overall, was the criterion, *Scientific Acceptability of Measure Properties*, met? **(*Reliability and Validity must be rated moderate or high*) Yes☐ No☐** **Provide rationale based on specific subcriteria:**
If the Committee votes No, STOP

3. Feasibility

Extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

3a. Byproduct of Care Processes

For clinical measures, the required data elements are routinely generated and used during care delivery (e.g., blood pressure, lab test, diagnosis, medication order).

3a.1. **Data Elements Generated as Byproduct of Care Processes.**

Generated or collected by and used by healthcare personnel during the provision of care (e.g., blood pressure, lab value, diagnosis, depression score), Coded by someone other than person obtaining original information (e.g., DRG, ICD-9 codes on claims), Abstracted from a record by someone other than person obtaining original information (e.g., chart abstraction for quality measure or registry)

If other:

3b. Electronic Sources

The required data elements are available in electronic health records or other electronic sources. If the required data are not in electronic health records or existing electronic sources, a credible, near-term path to electronic collection is specified.

3b.1. **To what extent are the specified data elements available electronically in defined fields?** *(i.e., data elements that are needed to compute the performance measure score are in defined, computer-readable fields)*

Some data elements are in defined fields in electronic sources

3b.2. **If ALL the data elements needed to compute the performance measure score are not from electronic sources, specify a credible, near-term path to electronic capture, OR provide a rationale for using other than electronic sources.**

To allow for widespread reporting across health plans and health care practices, this measure is collected through multiple data sources (administrative data, electronic clinical data, paper records). We anticipate as electronic health records become more widespread the reliance on paper record review will decrease.

3b.3. **If this is an eMeasure, provide a summary of the feasibility assessment in an attached file or make available at a measure-specific URL.**

Attachment:

3c. Data Collection Strategy

Demonstration that the data collection strategy (e.g., source, timing, frequency, sampling, patient confidentiality, costs associated with fees/licensing of proprietary measures) can be implemented (e.g., already in operational use, or testing demonstrates that it is ready to put into operational use). For eMeasures, a feasibility assessment addresses the data elements and measure logic and demonstrates the eMeasure can be implemented or feasibility concerns can be adequately addressed.

3c.1. **Describe what you have learned/modified as a result of testing and/or operational use of the measure regarding data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, time and cost of data collection, other feasibility/implementation issues.**

IF a PRO-PM, consider implications for both individuals providing PROM data (patients, service recipients, respondents) and those whose performance is being measured.

NCQA recognizes that, despite the clear specifications defined for HEDIS measures, data collection and calculation methods may vary, and other errors may affect the results. Thus, NCQA conducts an independent audit of all HEDIS collection and reporting processes, as well as an audit of the data which are manipulated by those processes, in order to verify that HEDIS specifications are met. NCQA has developed a precise, standardized methodology for verifying the integrity of HEDIS collection and calculation processes through a two-part program consisting of an overall information systems capabilities assessment followed by an evaluation of the MCO´s ability to comply with HEDIS specifications. NCQA-certified auditors using standard audit methodologies will help enable purchasers to make more reliable "apples-to-apples" comparisons between health plans.

The HEDIS Compliance Audit addresses the following functions:

1) information practices and control procedures

2) sampling methods and procedures

3) data integrity

4) compliance with HEDIS specifications

5) analytic file production

6) reporting and documentation

In addition to the HEDIS Audit, NCQA provides a system to allow "real-time" feedback from measure users. Our Policy Clarification Support System receives thousands of inquiries each year on over 100 measures. Through this system NCQA responds immediately to questions and identifies possible errors or inconsistencies in the implementation of the measure. This system is vital to the regular re-evaluation of NCQA measures.

Input from NCQA auditing and the Policy Clarification Support System informs the annual updating of all HEDIS measures including updating value sets and clarifying the specifications. Measures are re-evaluated on a periodic basis and when there is a significant change in evidence. During re-evaluation information from NCQA auditing and Policy Clarification Support System is used to inform evaluation of the scientific soundness and feasibility of the measure.

3c.2. Describe any fees, licensing, or other requirements to use any aspect of the measure as specified *(e.g., value/code set, risk model, programming code, algorithm)*.

Broad public use and dissemination of these measures is encouraged and NCQA has agreed with NQF that noncommercial uses do not require the consent of the measure developer. Use by health care physicians in connection with their own practices is not commercial use. Commercial use of a measure requires the prior written consent of NCQA. As used herein, "commercial use" refers to any sale, license or distribution of a measure for commercial gain, or incorporation of a measure into any product or service that is sold, licensed or distributed for commercial gain, even if there is no actual charge for inclusion of the measure.

4. Usability and Use

Extent to which potential audiences (e.g., consumers, purchasers, providers, policy makers) are using or could use performance results for both accountability and performance improvement to achieve the goal of high-quality, efficient healthcare for individuals or populations.

4a. Accountability and Transparency

Performance results are used in at least one accountability application within three years after initial endorsement and are publicly reported within six years after initial endorsement (or the data on performance results are available). If not in use at the time of initial endorsement, then a credible plan for implementation within the specified timeframes is provided.

4.1. Current and Planned Use

NQF-endorsed measures are expected to be used in at least one accountability application within 3 years and publicly reported within 6 years of initial endorsement in addition to performance improvement.

Planned	Current Use (for current use provide URL)
	Public Reporting NCQA Quality Compass NCQA State of Health Care Quality Report Medicaid Child Core Set http://www.ncqa.org/tabid/177/Default.aspx http://www.ncqa.org/tabid/836/Default.aspx https://www.medicaid.gov/Medicaid-CHIP-Program-Information/By-Topics/Quality-of-Care/CHIPRA-Initial-Core-Set-of-Childrens-Health-Care-Quality-Measures.html Quality Improvement with Benchmarking (external benchmarking to multiple organizations) NCQA Quality Compass NCQA State of Health Care Quality Report Medicaid Child Core Set http://www.ncqa.org/tabid/177/Default.aspx http://www.ncqa.org/tabid/836/Default.aspx https://www.medicaid.gov/Medicaid-CHIP-Program-Information/By-Topics/Quality-of-Care/CHIPRA-Initial-Core-Set-of-Childrens-Health-Care-Quality-Measures.html

4a.1. For each CURRENT use, checked above, provide:

- Name of program and sponsor
- Purpose
- Geographic area and number and percentage of accountable entities and patients included

QUALITY COMPASS: This measure is used in Quality Compass which is an indispensable tool used for selecting a health plan, conducting competitor analysis, examining quality improvement and benchmarking plan performance. Provided in this tool is the ability to generate custom reports by selecting plans, measures, and benchmarks (averages and percentiles) for up to three trended years. Results in table and graph formats offer simple comparison of plans' performance against competitors or benchmarks.

STATE OF HEALTH CARE QUALITY REPORT: This measure is publically reported nationally and by geographic regions in the NCQA State of Health Care annual report. This annual report published by NCQA summarizes findings on quality of care. In 2012 the report included measures on 11.5 million Medicare Advantage beneficiaries in 455 Medicare Advantage health plans, 99.4 million members in 404 commercial health plans, and 14.3 million Medicaid beneficiaries in 136 plans across 50 states.

MEDICAID CHILD CORE SET: These are a core set of health quality measures for Medicaid-enrolled children. The Medicaid Child Core Set was identified by the Centers for Medicare & Medicaid (CMS) in partnership with the Agency for HealthCare Research and Quality (AHRQ). The data collected from these measures will help CMS to better understand the quality of health care that children enrolled in Medicaid receive nationally. On December 29, 2009, the Secretary posted for public comment in the Federal Register, an initial core set of 24 children´s health care quality measures for voluntary use by Medicaid and CHIP programs. The CHIPRA

legislation provides that the Secretary shall issue updates to the Child Core Set beginning in January 2013 and annually thereafter. CMS worked with the National Quality Forum's (NQF) Measures Application Partnership (MAP) to review the Child Core Set and to identify ways to improve it. State data derived from the core measures are part of the Secretary´s Annual Report on the Quality of Care for Children in Medicaid and CHIP.

4a.2. **If not currently publicly reported OR used in at least one other accountability application (e.g., payment program, certification, licensing) what are the reasons?** (*e.g., Do policies or actions of the developer/steward or accountable entities restrict access to performance results or impede implementation?*)
NA

4a.3. **If not currently publicly reported OR used in at least one other accountability application, provide a credible plan for implementation within the expected timeframes -- any accountability application within 3 years and publicly reported within 6 years of initial endorsement.** (*Credible plan includes the specific program, purpose, intended audience, and timeline for implementing the measure within the specified timeframes. A plan for accountability applications addresses mechanisms for data aggregation and reporting.*)
NA

4b. Improvement

Progress toward achieving the goal of high-quality, efficient healthcare for individuals or populations is demonstrated. If not in use for performance improvement at the time of initial endorsement, then a credible rationale describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.

4b.1. **Progress on Improvement. (Not required for initial endorsement unless available.)**
Performance results on this measure (current and over time) should be provided in 1b.2 and 1b.4. Discuss:

- **Progress (trends in performance results, number and percentage of people receiving high-quality healthcare)**
- **Geographic area and number and percentage of accountable entities and patients included**

Performance results (listed in 1b.2 and 1b.4) show that rates have been steady over the past three years among Medicaid plans. However, there continues to be variation in the Frequency of Prenatal Care measure. In 2015, for the less than 21 percent bracket, there was a 29 percentage point difference between plans in the 10th percentile vs the 90th percentile. In 2015, for the 81+ percent, there was a 48 percentage point difference between plans in the 10th percentile vs the 90th percentile. These gaps in performance underscore the opportunity for improvement.

4b.2. **If no improvement was demonstrated, what are the reasons? If not in use for performance improvement at the time of initial endorsement, provide a credible rationale that describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.**
It is not clear why improvement has not occurred, though as noted there continues to be variation among health care plans. The measure is actively used in programs for both health plan and state reporting. The measure assesses pregnant women's access to care; it is possible that health insurance expansion will increase access to prenatal care over time.

4c. Unintended Consequences

The benefits of the performance measure in facilitating progress toward achieving high-quality, efficient healthcare for individuals or populations outweigh evidence of unintended negative consequences to individuals or populations (if such evidence exists).

4c.1. **Were any unintended negative consequences to individuals or populations identified during testing; OR has evidence of unintended negative consequences to individuals or populations been reported since implementation? If so, identify the negative unintended consequences and describe how benefits outweigh them or actions taken to mitigate them.**
There were no identified unintended consequences for this measure during testing or since implementation.

5. Comparison to Related or Competing Measures

If a measure meets the above criteria and there are endorsed or new related measures (either the same measure focus or the same target population) or competing measures (both the same measure focus and the same target population), the measures are compared to address harmonization and/or selection of the best measure.

5. Relation to Other NQF-endorsed Measures
Are there related measures (conceptually, either same measure focus or target population) or competing measures (conceptually both the same measure focus and same target population)? If yes, list the NQF # and title of all related and/or competing measures.
Yes

5.1a. **List of related or competing measures (selected from NQF-endorsed measures)**
1517 : Prenatal & Postpartum Care (PPC)

5.1b. **If related or competing measures are not NQF endorsed please indicate measure title and steward.**

5a. Harmonization

The measure specifications are harmonized with related measures;
OR
The differences in specifications are justified

5a.1. **If this measure conceptually addresses EITHER the same measure focus OR the same target population as NQF-endorsed measure(s):**
Are the measure specifications completely harmonized?
Yes

5a.2. **If the measure specifications are not completely harmonized, identify the differences, rationale, and impact on interpretability and data collection burden.**

5b. Competing Measures

The measure is superior to competing measures (e.g., is a more valid or efficient way to measure);
OR
Multiple measures are justified.

5b.1. **If this measure conceptually addresses both the same measure focus and the same target population as NQF-endorsed measure(s):**
Describe why this measure is superior to competing measures (e.g., a more valid or efficient way to measure quality); OR provide a rationale for the additive value of endorsing an additional measure. (Provide analyses when possible.)

Appendix

A.1 Supplemental materials may be provided in an appendix. All supplemental materials (such as data collection instrument or methodology reports) should be organized in one file with a table of contents or bookmarks. If material pertains to a specific submission form number, that should be indicated. Requested information should be provided in the submission form and required attachments. There is no guarantee that supplemental materials will be reviewed.

Attachment:

Contact Information

Co.1 **Measure Steward (Intellectual Property Owner):** National Committee for Quality Assurance
Co.2 **Point of Contact:** Bob, Rehm, nqf@ncqa.org, 202-955-1728-
Co.3 **Measure Developer if different from Measure Steward:** National Committee for Quality Assurance
Co.4 **Point of Contact:** Bob, Rehm, nqf@ncqa.org, 202-955-1728-

Additional Information

Ad.1 **Workgroup/Expert Panel involved in measure development**
Provide a list of sponsoring organizations and workgroup/panel members' names and organizations. Describe the members' role in measure development.
Perinatal Care Measurement Advisory Panel

Eli Y. Adashi, MD, MS, CPE, FACOG, Brown University

Hani Atrash, MD, Health Resources and Services Administration

Sean Currigan, MPH, American College of Obstetricians and Gynecologist

Nicole Garro, MPH, March of Dimes

Rebekah E. Gee, MD, MPH, FACOG, Louisiana State University School of Public Health

Tina D. Groat, MD, MBA, FACOG, UnitedHealthcare

Renee Miskimmin, MD, MBA, FAAFP, Virginia Premier Health Plan

Catherine Ruhl, MS, CNM, Association of Women's Health, Obstetric and Neonatal Nurses

Carol Sakala, PhD, MSPH, National Partnership for Woman & Families

Carolyn L. Westhoff, MD, MSc, Planned Parenthood Federation of America

Measure Developer/Steward Updates and Ongoing Maintenance
Ad.2 **Year the measure was first released:** 1997
Ad.3 **Month and Year of most recent revision:** 07, 2014
Ad.4 **What is your frequency for review/update of this measure?** Approximately every 3 years, sooner if the clinical guidelines have changed significantly.
Ad.5 **When is the next scheduled review/update for this measure?** 12, 2017

Ad.7 **Disclaimers:** These performance measures are not clinical guidelines and do not establish a standard of medical care, and have not been tested for all potential applications.

THE MEASURES AND SPECIFICATIONS ARE PROVIDED "AS IS" WITHOUT WARRANTY OF ANY KIND.

Ad.8 **Additional Information/Comments:** NCQA Notice of Use. Broad public use and dissemination of these measures is encouraged and NCQA has agreed with NQF that noncommercial uses do not require the consent of the measure developer. Use by health care physicians in connection with their own practices is not commercial use. Commercial use of a measure requires the prior written consent of NCQA. As used herein, "commercial use" refers to any sale, license or distribution of a measure for commercial gain, or incorporation of a measure into any product or service that is sold, licensed or distributed for commercial gain, even if there is no actual charge for inclusion of the measure.

These performance measures were developed and are owned by NCQA. They are not clinical guidelines and do not establish a standard of medical care. NCQA makes no representations, warranties or endorsement about the quality of any organization or physician that uses or reports performance measures, and NCQA has no liability to anyone who relies on such measures. NCQA holds a copyright in these measures and can rescind or alter these measures at any time. Users of the measures shall not have the right to alter, enhance or otherwise modify the measures, and shall not disassemble, recompile or reverse engineer the source code or object code relating to the measures. Anyone desiring to use or reproduce the measures without modification for a noncommercial purpose may do so without obtaining approval from NCQA. All commercial uses must be approved by NCQA and are subject to a license at the discretion of NCQA.

MEASURE WORKSHEET

This document summarizes the evaluation of the measure as it progresses through NQF's Consensus Development Process (CDP). The information submitted by measure developers/stewards is included after the Brief Measure Information, Preliminary Analysis, and Pre-meeting Public and Member Comments sections.

To navigate the links in the worksheet: Ctrl + click link to go to the link; ALT + LEFT ARROW to return

Brief Measure Information

NQF #: 1517
Measure Title: Prenatal & Postpartum Care (PPC)
Measure Steward: National Committee for Quality Assurance
Brief Description of Measure: The percentage of deliveries of live births between November 6 of the year prior to the measurement year and November 5 of the measurement year. For these women, the measure assesses the following facets of prenatal and postpartum care:
Rate 1: Timeliness of Prenatal Care. The percentage of deliveries that received a prenatal care visit as a member of the organization in the first trimester or within 42 days of enrollment in the organization.
Rate 2: Postpartum Care. The percentage of deliveries that had a postpartum visit on or between 21 and 56 days after delivery.
Developer Rationale: This measure assesses appropriate prenatal and postpartum care by seeking to ensure that pregnant women receive both timely prenatal care and postpartum visits. Women who utilize prenatal care can minimize their risk for pregnancy complications and negative birth outcomes, which include higher risk of infant morbidity and mortality. Postpartum care is an opportunity to assess the well-being of the mother post-delivery and to provide important services such as maternal depression screening. Adherence to this measure could improve access to services and provide opportunities to improve health of both the infant and the mother.

Numerator Statement: This measure assesses whether pregnant women had timely prenatal and postpartum care visits. It has two rates, one assessing the timeliness of prenatal visits, and one assessing the timeliness of postpartum visits.
Denominator Statement: The percentage of deliveries of live births between November 6 of the year prior to the measurement year and November 5 of the measurement year.
Denominator Exclusions: Non-live births

Measure Type: Process
Data Source: Administrative claims, Electronic Clinical Data, Paper Medical Records
Level of Analysis: Health Plan, Integrated Delivery System

IF Endorsement Maintenance – Original Endorsement Date: Aug 15, 2011 **Most Recent Endorsement Date:** Aug 15, 2011

Maintenance of Endorsement -- Preliminary Analysis

To maintain NQF endorsement endorsed measures are evaluated periodically to ensure that the measures still meets the NQF endorsement criteria ("maintenance"). The emphasis for maintaining endorsement is focused on how effective the measure is for promoting improvements in quality. Endorsed measures should have some experience from the field to inform the evaluation. The emphasis for maintaining endorsement is noted for each criterion.

Criteria 1: Importance to Measure and Report

1a. Evidence
Maintenance measures – less emphasis on evidence unless there is new information or change in evidence since the prior evaluation.

1a. Evidence. The evidence requirements for a *process or intermediate outcome* measure is that it is based on a

systematic review (SR) and grading of the body of empirical evidence where the specific focus of the evidence matches what is being measured.

The developer provides the following evidence for this measure:

- **Systematic Review of the evidence specific to this measure?** ☐ **Yes** ☒ **No**
- **Quality, Quantity and Consistency of evidence provided?** ☐ **Yes** ☒ **No**
- **Evidence graded?** ☐ **Yes** ☒ **No**

Summary of prior review in 2010

The prior Committee noted that this is measure only assesses visits but not the content of those visits. The Committee agreed that ACOG guidelines recommend a schedule of prenatal visits that are based primarily on expert consensus. The Committee acknowledged that data does show that patients who have no prenatal care have worse outcomes.

Changes to evidence from last review

☒ **The developer attests that there have been no changes in the evidence since the measure was last evaluated.**

☐ **The developer provided updated evidence for this measure:**

Updates: none

Exception to evidence

Since the evidence for this measure is expert consensus rather than empirical evidence, it is insufficient to meet NQF's criterion for evidence. However, an exception to the evidence criterion is allowed if the Committee agrees that empirical evidence is not needed to hold providers accountable for the measure.

Guidance from the Evidence Algorithm

Process measure (Box 1) --> not based on a SR (Box 3) --> no empirical evidence (Box 7) --> systematic assessment of expert opinion (Box 11) --> if Committee agrees it is OK/beneficial to hold providers accountable for performance in the absence of empirical evidence of benefits to patients → rate as INSUFFICIENT WITH EXCEPTION

Staff NOTE: The Committee will first vote on the criterion. If a > 60% votes for insufficient then a second vote will be taken to determine whether the Committee wishes to pass the evidence criterion with exception.

Questions for the Committee:

- *For possible exception to the evidence criterion:*
  - *Are there, or could there be, performance measures of a related health outcome, OR evidence-based intermediate clinical outcomes, intervention/treatment?*
  - *Is there evidence of a systematic assessment of expert opinion beyond those involved in developing the measure?*
  - *Does the SC agree that it is acceptable (or beneficial) to hold providers accountable without empirical evidence?*

Preliminary rating for evidence: ☐ **High** ☐ **Moderate** ☐ **Low** ☒ **Insufficient**

1b. Gap in Care/Opportunity for Improvement and 1b. Disparities
Maintenance measures – increased emphasis on gap and variation

1b. Performance Gap. The performance gap requirements include demonstrating quality problems and opportunity for improvement.

The developer provides the following data on performance:

	2015	2014	2013
Timeliness of prenatal care			
Commercial plans (SD)	84% (13%)	87% (13%)	85% (15%)
10-90 percentile	67 -95%	70 – 97%	70 – 96%
Medicaid plans (SD)	82% (8.9%)	82% (12%)	83% (12%)
10-90 percentile	69 – 92%	70 – 93%	71 – 93%
Postpartum care			
Commercial plans (SD)	73% (15%)	76% (15%)	75% (15%)
10-90 percentile	46 – 88%	48 – 89%	49 – 89%
Medicaid plans (SD)	62% (9.3%)	61% (11%)	63% (9.8%)
10-90 percentile	49 – 72%	48 – 74%	51 – 74%

The 2010 evaluation provided the following national HEDIS data:

	2006	2007
Timeliness of prenatal care	81.24	81.37
Postpartum care	59.08	58.6

Disparities

No additional data addressing disparities is presented.

Questions for the Committee:

- *This measure has been endorsed by NQF for 5 years. How has performance changed over time?*
- *Is there a gap in care that warrants a national performance measure?*
- *Can this measure be used to address disparities?*

Preliminary rating for opportunity for improvement: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 1: Importance to Measure and Report (including 1a, 1b, 1c)

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability

2a1. Reliability Specifications

Maintenance measures – no change in emphasis – specifications should be evaluated the same as with new measures

2a1. Specifications requires the measure, as specified, to produce consistent (reliable) and credible (valid) results about the quality of care when implemented.

Data source(s): Administrative claims and medical records

Specifications:

- Two sets of specifications are presented – one for administrative claims and one for medical records
  - The admin specifications capture a variety of prenatal visit codes
  - The medical record specifications include "evidence of one of the following" – OB history, physical

examination, a list of laboratory tests
- The denominator is "deliveries of live births" in the previous year. Non-live births are excluded.
- The spreadsheet containing the value sets include codes for both ICD-9 and ICD-10.
- The developers report the following changes to the specifications:
  - Deleted the use of infant claims to identify deliveries.
  - Clarified the tests that must be included to meet criteria for an obstetric panel in the medical record specification. These are as follows: hematocrit, differential WBC count, platelet count, hepatitis B surface antigen, rubella antibody, syphilis test, RBC antibody screen, Rh and ABO blood typing.
- A calculation algorithm is included.

Questions for the Committee :
- *Are all the data elements clearly defined? Are all appropriate codes included?*
- *Is it appropriate to exclude "non-live births"?*
- *Is the logic or calculation algorithm clear?*
- *Since there are two sets of specifications for different data sources, is it likely this measure consistently implemented and the results are comparable?*

2a2. Reliability Testing Testing attachment
Maintenance measures – less emphasis if no new testing data provided

2a2. Reliability testing demonstrates if the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise enough to distinguish differences in performance across providers.

For maintenance measures, summarize the reliability testing from the prior review (2010):
Measure score reliability was demonstrated through a signal to noise analysis (beta binomial method) for 2010 data:

Rate - Timeliness of Prenatal Care:
Reliability - Commercial Plans: 0.9961 Reliability - Medicaid Plans: 0.9564

Rate - Postpartum Care:
Reliability - Commercial Plans: 0.9944 Reliability - Medicaid Plans: 0.9217

This is a commonly used method to assess reliability. The higher the reliability score, the greater is the confidence with which one can distinguish the performance of one plan from another. A reliability score greater than or equal to 0.7 is generally considered acceptable.

Describe any updates to testing none

SUMMARY OF TESTING
Reliability testing level ☒ **Measure score** ☐ **Data element** ☐ **Both**
Reliability testing performed with the data source and level of analysis indicated for this measure ☐ **Yes** ☐ **No**

Method(s) of reliability testing see above

Results of reliability testing see above

Guidance from the Reliability Algorithm
Precise specifications (Box 1) [may be an issue] → empirical reliability testing (Box 2) → measure score testing (Box 4) → appropriate method (Box 5) → High certainty or confidence that the performance measure scores are reliable (Box 6a)

Questions for the Committee:

- *The developer attests there is no new testing data since the last NQF endorsement review. Does the Committee agree there is no need for repeat discussion and voting on reliability?*

Preliminary rating for reliability: ☒ High ☐ Moderate ☐ Low ☐ Insufficient
High IF no concerns about two sets of specifications

2b. Validity
Maintenance measures – less emphasis if no new testing data provided

2b1. Validity: Specifications

2b1. Validity Specifications. This section should determine if the measure specifications are consistent with the evidence.

Specifications consistent with evidence in 1a. ☒ Yes ☐ Somewhat ☐ No
Specification not completely consistent with evidence

Question for the Committee:

- *Are the specifications consistent with the evidence?*

2b2. Validity testing

2b2. Validity Testing should demonstrate the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality.

For maintenance measures, summarize the validity testing from the prior review: Face validity assessment only by a panel of stakeholders and experts

Describe any updates to validity testing none

SUMMARY OF TESTING
Validity testing level ☒ Measure score ☐ Data element testing against a gold standard ☐ Both

Method of validity testing of the measure score:

- ☒ **Face validity only**
- ☐ **Empirical validity testing of the measure score**

Validity testing method: Face validity only

Validity testing results: "This measure was deemed valid by the expert panel."

Questions for the Committee:

- *The previous Committee raised concerns about the lack of specificity for services that should be provided at the visit. Does a visit-based measure provide valid information about the quality of care provided to patients?*
- *Is the test sample adequate to generalize for widespread implementation?*
- *Do the results demonstrate sufficient validity so that conclusions about quality can be made?*
- *Do you agree that the score from this measure as specified is an indicator of quality?*

2b3-2b7. Threats to Validity

2b3. Exclusions: No information is provided.

- The only exclusion is non-live births. No data is provided on the frequency of this exclusion.

Questions for the Committee:

- *Are the exclusions consistent with the evidence?*
- *Should "non-live births" be excluded?*
- *Are any patients or patient groups inappropriately excluded from the measure?*
- *Are the exclusions/exceptions of sufficient frequency and variation across providers to be needed (and outweigh the data collection burden)?*

2b4. Risk adjustment: **Risk-adjustment method** ☒ **None** ☐ **Statistical model** ☐ **Stratification**

2b5. Meaningful difference *(can statistically significant and clinically/practically meaningful differences in performance measure scores can be identified)*:

- The developer reports that meaningful differences are determined by "Comparison of means and percentiles; analysis of variance against established benchmarks; if sample size is >400, we would use an analysis of variance."
- See data table in 1b.

Question for the Committee:

- *Does this measure identify meaningful differences about quality?*

2b6. Comparability of data sources/methods: The developer did not respond to the question.

- Two sets of specifications are presented. No information on comparability is discussed.
- No information on the frequency of use of each data source is provided.

Question for the Committee:

- *Are the two sets of specifications likely to produce comparable results?*

2b7. Missing Data The developer reports "NA" for this question.

Guidance from the algorithm:

Consistent with the evidence (Box 1) → Potential threats to validity addressed (Box 2)[comparability of data sources may be an issue] →empirical validity testing (Box 4) → face validity systematically assessed (Box 4) → Highest rating possible is MODERATE if threats to validity adequately addressed

Question for the Committee:

- ***The developer attests there is no new testing for validity since the last NQF endorsement review. Does the Committee agree there is no need for repeat discussion and voting on validity?***

Preliminary rating for validity: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

IF comparability of data sources is not an issue. In the absence of empirical validity testing, the highest rating possible is modereate.

Committee pre-evaluation comments

Criteria 2: Scientific Acceptability of Measure Properties (including all 2a, 2b, and 2d)

Criterion 3. Feasibility
Maintenance measures – no change in emphasis – implementation issues may be more prominent

3. Feasibility is the extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

- The developer states that "To allow for widespread reporting across health plans and health care practices, this measure is collected through multiple data sources (administrative data, electronic clinical data, paper records)."
- Measures collected vis administrative claims are quite feasible. The burden of paper medical record review is considerable.

Questions for the Committee:

- *This measure is specified for health plans and integrated delivery systems. Are there any of these entities that would not be collecting this data through claims?*
- *Does the Committee have any concerns about the feasibility of this measure?*

Preliminary rating for feasibility: ☒ **High** ☐ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 3: Feasibility

Criterion 4: Usability and Use
Maintenance measures – increased emphasis – much greater focus on measure use and usefulness, including both impact /improvement and unintended consequences

4. Usability and Use evaluate the extent to which audiences (e.g., consumers, purchasers, providers, policymakers) use or could use performance results for both accountability and performance improvement activities.

Current uses of the measure
Publicly reported? ☒ **Yes** ☐ **No**

Current use in an accountability program? ☒ **Yes** ☐ **No**
OR
Planned use in an accountability program? ☐ **Yes** ☐ **No**

Accountability program details
The developer reports on multiple uses of the measures including public reporting and quality improvement (i.e., NCQA Quality Compass, NCQA State of Health Care Quality Report, Medicaid Child Core Set). The developer provided the following information for each program.

- QUALITY COMPASS: http://www.ncqa.org/tabid/177/Default.aspx This measure is used in Quality Compass which is an indispensable tool used for selecting a health plan, conducting competitor analysis, examining quality improvement and benchmarking plan performance. Provided in this tool is the ability to generate custom reports by selecting plans, measures, and benchmarks (averages and percentiles) for up to three trended years. Results in table and graph formats offer simple comparison of plans' performance against competitors or benchmarks.
- STATE OF HEALTH CARE QUALITY REPORT: This measure is publically reported nationally and by geographic regions in the NCQA State of Health Care annual report. This annual report published by NCQA summarizes findings on quality of care. In 2012 the report included measures on 11.5 million Medicare Advantage beneficiaries in 455 Medicare Advantage health plans, 99.4 million members in 404 commercial health plans, and 14.3 million Medicaid beneficiaries in 136 plans across 50 states.
- MEDICAID CHILD CORE SET: http://www.ncqa.org/tabid/836/Default.aspx These are a core set of health quality measures for Medicaid-enrolled children. The Medicaid Child Core Set was identified by the Centers for Medicare & Medicaid (CMS) in partnership with the Agency for HealthCare Research and Quality (AHRQ). The

data collected from these measures will help CMS to better understand the quality of health care that children enrolled in Medicaid receive nationally. On December 29, 2009, the Secretary posted for public comment in the Federal Register, an initial core set of 24 children´s health care quality measures for voluntary use by Medicaid and CHIP programs. The CHIPRA legislation provides that the Secretary shall issue updates to the Child Core Set beginning in January 2013 and annually thereafter. CMS worked with the National Quality Forum's (NQF) Measures Application Partnership (MAP) to review the Child Core Set and to identify ways to improve it. State data derived from the core measures are part of the Secretary´s Annual Report on the Quality of Care for Children in Medicaid and CHIP.

Improvement results See data table in 1b.

The developer comments: "Performance results show that rates have been steady over the past three years among commercial and Medicaid plans." "It is not clear why improvement has not occurred, though as noted there continues to be variation among health care plans. The measure is actively used in programs for both health plan and state reporting."

Unexpected findings (positive or negative) during implementation

The developer reports : " There were no unexpected finding or benefits"

Potential harms not discussed

Feedback :

- In 2015 MAP Medicaid Task Force supported the continued use of this measure in the Medicaid Child Core Set.

***Questions for the Committee**:*

- *Committee members should share any experience with use of this measure in their workplace.*
- *How do health plans and systems use this measure?*
- *This measure has been in use for many years – what has been the impact?*
- *Is the Committee aware of any unintended consequences or potential harms from the measure? Any unexpected benefits from use of the measure?*
- *Do the benefits of the measure outweigh any potential unintended consequences?*

Preliminary rating for usability and use: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 4: Usability and Use

Criterion 5: Related and Competing Measures

Related or competing measures

Related measure: 1391 : Frequency of Ongoing Prenatal Care (FPC)

Harmonization - both measures are from NCQA and are harmonized for use together.

Pre-meeting public and member comments

-

NQF #: 1517 **NQF Project:** Child Health Quality Measures 2010

1. IMPACT, OPPORTUITY, EVIDENCE - IMPORTANCE TO MEASURE AND REPORT

Importance to Measure and Report is a threshold criterion that must be met in order to recommend a measure for endorsement. All three subcriteria must be met to pass this criterion. See guidance on evidence.

Measures must be judged to be important to measure and report in order to be evaluated against the remaining criteria. **(evaluation criteria)**

1c.1 **Structure-Process-Outcome Relationship** *(Briefly state the measure focus, e.g., health outcome, intermediate clinical outcome, process, structure; then identify the appropriate links, e.g., structure-process-health outcome; process-health outcome; intermediate clinical outcome-health outcome)***:**

Proper perinatal care is associated with improved birth outcomes. For example, one study found that 25.6 percent of women who did not receive prenatal care delivered preterm infants compared to 9.2 percent of women who received even a minimum amount of prenatal care (Vintzileos et al., 2002).

In 2001, infants of mothers who received no prenatal care had an infant mortality rate of 34.8 per 1,000 live births, compared to an infant mortality rate of only 6.2 per 1,000 when prenatal care was initiated in the first trimester of pregnancy (Matthews et al., 2003). Observational studies have consistently shown that groups having more post-delivery visits have lower maternal, fetal and neonatal illness and mortality.

Regarding postpartum visits, not only do many women experience some degree of emotional liability in the postpartum period, which warrants a follow-up visit, but they will also need personalized care during this time to hasten the development of a healthy mother-infant relationship and a sense of maternal confidence (ACOG, 2002). Should the pregnancy have an abnormal outcome, the postpartum visit is an advantageous time to discuss implications of such conditions as diabetes mellitus, intrauterine growth restriction, preterm birth, hypertension or other conditions that may recur in any future pregnancies (ACOG, 2002). The postpartum visit is also an ideal time to begin preconceptional counseling for patients who may wish to have future pregnancies.

1c.2-3 **Type of Evidence** *(Check all that apply)***:**

1c.4 **Directness of Evidence to the Specified Measure** *(State the central topic, population, and outcomes addressed in the body of evidence and identify any differences from the measure focus and measure target population)***:**

The goal of the prenatal contact is to exchange information and identify existing risk factors that may impact the pregnancy (DoD/VA, 2002). Lack of prenatal care can be considered a high-risk factor for postneonatal death. A study that sought to determine the association between prenatal care (defined as one visit) and postneonatal death rates (defined as the number of deaths of infants between 28 and 365 days of life) found that the postneonatal deaths among women who had prenatal care was 2.1 per 1,000 women, whereas the rate among women without prenatal care was 5.9 per 1,000 (Vintzileos, A, et al., 2002). These rates applied to women without high-risk conditions. Women whose prenatal care fails to meet established standards are at a greater risk for pregnancy complications and negative birth outcomes (National Center for Health Statistics, 1997).

The goal of postpartum care is to assess the physical and psychosocial status of the mother after the mother's discharge. The majority of maternal and neonatal deaths, as well as a significant burden of long-term morbidity, occur during the postpartum period (WHO, 1998). The postpartum visit should include obtaining an interval history and performing a physical exam to evaluate the patient's current status. Additionally, the emotional status of a woman whose pregnancy had an abnormal outcome should be reviewed, as many women experience some degree of emotional liability during the postpartum period. It is also an advisable time to begin preconceptional counseling for patients who may wish to have future pregnancies (ACOG Guidelines, 2002).

1c.5 Quantity of Studies in the Body of Evidence *(Total number of studies, not articles)***:**

1c.6 Quality of Body of Evidence (*Summarize the certainty or confidence in the estimates of benefits and harms to patients across studies in the body of evidence resulting from study factors. Please address: a) study design/flaws; b) directness/indirectness of the evidence to this measure (e.g., interventions, comparisons, outcomes assessed, population included in the evidence); and c) imprecision/wide confidence intervals due to few patients or events):*

1c.7 Consistency of Results across Studies *(Summarize the consistency of the magnitude and direction of the effect):*

1c.8 Net Benefit *(Provide estimates of effect for benefit/outcome; identify harms addressed and estimates of effect; and net benefit - benefit over harms)***:**

1c.9 Grading of Strength/Quality of the Body of Evidence. Has the body of evidence been graded?

1c.10 If body of evidence graded, identify the entity that graded the evidence including balance of representation and any disclosures regarding bias:

1c.11 System Used for Grading the Body of Evidence: Expert Consensus

1c.12 If other, identify and describe the grading scale with definitions:

1c.13 Grade Assigned to the Body of Evidence: Good

1c.14 Summary of Controversy/Contradictory Evidence: None

1c.15 Citations for Evidence other than Guidelines*(Guidelines addressed below)*:

Veterans Health Administration, Department of Defense. DoD/VA clinical practice guideline for the management of uncomplicated pregnancy. Washington (DC): Department of Veterans Affairs; 2002 October.

Department of Reproductive Health and Research (RHR), World Health Organization (WHO). Postpartum care of the mother and newborn: a practical guide; 1998.

Vintzileos, A, Ananth, C, Smulian, JC, Scorza, WE, Knuppel, RA. The impact of prenatal care on postneonatal deaths in the presence and absence of antenatal high-risk conditions. Am J Obstet Gynecol November 2002; 187(5):1258-1262

1c.16 Quote verbatim, <u>the specific guideline recommendation</u> *(Including guideline # and/or page #)***:**

ACOG

Timeliness of Prenatal Care: Once pregnancy occurs, patients should have early contact with an obstetrician to begin counseling about prenatal testing and to develop a management plan. The frequency of subsequent antepartum office visits is determined by the individual needs of the woman and the assessment of her risks. According to ACOG guidelines, for an uncomplicated pregnancy, the ACOG guidelines suggest the following frequency of office visits: monthly office visits from the initial prenatal visit until 29 weeks of pregnancy; weekly office visits from 36 weeks until delivery; office visits every two to three weeks from 29 weeks to 36 weeks of pregnancy. (ACOG guidelines, 2010)

Postpartum Care: The timing of the postpartum visit has been a topic for debate. According to the ACOG, the mother should visit her physician for a postpartum review and examination approximately 4 to 6 weeks after delivery. This interval may be modified according to the needs of the patient with medical, obstetric, or intercurrent complications (ACOG Guidelines, 2002).

The DoD/VA clinical practice guideline for management of uncomplicated pregnancy

Spports the recommended 8 weeks after delivery postpartum visit. Evidence suggests that eight weeks is the optimal time to decrease the rate of false positive cervical smears, though consideration of the mother's schedule should also be taken into account (2002).

A visit within 7-14 days of delivery may be advisable after a cesarean delivery or a complicated gestation, primarily to assess the surgical wounds and healing. The standard postpartum care visit is recommended in follow-up to this initial visit (ACOG Guidelines) to ensure the woman's uterus has reduced to its normal size, and to conduct a depression screening and family planning counseling.

1c.17 Clinical Practice Guideline Citation: American Academy of Pediatrics and The American College of Obstetricians and Gynecologists. Guidelines for Perinatal Care (5th Edition). October 2002.

1c.18 National Guideline Clearinghouse or other URL: Routine prenatal and postnatal care. http://www.guideline.gov/content.aspx?id=13174&search=prenatal+and+postpartum+care

1c.19 Grading of Strength of Guideline Recommendation. Has the recommendation been graded?

1c.20 If guideline recommendation graded, identify the entity that graded the evidence including balance of representation and any disclosures regarding bias:

1c.21 System Used for Grading the Strength of Guideline Recommendation: USPSTF

1c.22 If other, identify and describe the grading scale with definitions:

1c.23 Grade Assigned to the Recommendation: Good

1c.24 Rationale for Using this Guideline Over Others: The measures are access and use of service measures that are based on the body of evidence and guidelines regarding perinatal care.

Based on the NQF descriptions for rating the evidence, what was the <u>developer's assessment</u> of the quantity, quality, and consistency of the body of evidence?

1c.25 Quantity: **1c.26** Quality: **1c.27** Consistency**:**

1. Evidence, Performance Gap, Priority – Importance to Measure and Report

Extent to which the specific measure focus is evidence-based, important to making significant gains in healthcare quality, and improving health outcomes for a specific high-priority (high-impact) aspect of healthcare where there is variation in or overall less-than-optimal performance. ***Measures must be judged to meet all subcriteria to pass this criterion and be evaluated against the remaining criteria.***

1a. Evidence to Support the Measure Focus – See attached Evidence Submission Form

1517_Evidence_MSF5.0_Data-635278463206876303-635920201745982874.doc

1b. Performance Gap

Demonstration of quality problems and opportunity for improvement, i.e., data demonstrating:

- considerable variation, or overall less-than-optimal performance, in the quality of care across providers; and/or
- disparities in care across population groups.

1b.1. **Briefly explain the rationale for this measure** *(e.g., the benefits or improvements in quality envisioned by use of this measure)*

This measure assesses appropriate prenatal and postpartum care by seeking to ensure that pregnant women receive both timely prenatal care and postpartum visits. Women who utilize prenatal care can minimize their risk for pregnancy complications and negative birth outcomes, which include higher risk of infant morbidity and mortality. Postpartum care is an opportunity to assess the well-being of the mother post-delivery and to provide important services such as maternal depression screening. Adherence to this measure could improve access to services and provide opportunities to improve health of both the infant and the mother.

1b.2. **Provide performance scores on the measure as specified (current and over time) at the specified level of analysis.** *(This is required for endorsement maintenance. Include mean, std dev, min, max, interquartile range, scores by decile. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included). This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*

The following data are extracted from HEDIS data collection reflecting the most recent years of measurement for this measure. Performance data are at the health plan level and summarized by mean, standard deviation, minimum, maximum, and performance at the 10th, 25th, 50th, 75th and 90th percentile. Data are stratified by year and product line (i.e. commercial and Medicaid).

Rate 1: Timeliness of Prenatal Care

Commercial Rate

YEAR| MEAN | ST DEV | 10TH | 25TH | 50TH | 75TH | 90TH | Interquartile Range
2015 | 84% | 13% | 67% | 80% | 87% | 92% | 95% | 12.3%
2014 | 87% | 13% | 70% | 85% | 91% | 95% | 97% | 9.8%
2013 | 85% | 15% | 70% | 83% | 90% | 94% | 96% | 11.8%

Medicaid Rate

YEAR| MEAN | ST DEV | 10TH | 25TH | 50TH | 75TH | 90TH | Interquartile Range
2015 | 82% | 8.9% | 69% | 78% | 85% | 89% | 92% |11.1%
2014 | 82% | 12% | 70% | 78% | 84% | 89% | 93% | 11.5%
2013 | 83% | 12% | 71% | 80% | 86% | 90% | 93% | 9.7%

Rate 2: Postpartum Care

Commercial Rate

YEAR| MEAN | ST DEV | 10TH | 25TH | 50TH | 75TH | 90TH | Interquartile Range
2015 | 73% | 15% | 46% | 69% | 76% | 84% | 88% | 15.1%
2014 | 76% | 15% | 48% | 73% | 80% | 86% | 89% | 13.1%

2013 | 75% | 15% | 49% | 72% | 79% | 85% | 89% | 12.8%

Medicaid Rate
YEAR| MEAN | ST DEV | 10TH | 25TH | 50TH | 75TH | 90TH | Interquartile Range
2015 | 62% | 9.3% | 49% | 56% | 63% | 69% | 72% | 13.2%
2014 | 61% | 11% | 48% | 56% | 63% | 69% | 74% | 13.2%
2013 | 63% | 9.8% | 51% | 58% | 64% | 70% | 74% | 11.9%

Data for this measure are from the Healthcare Effectiveness Data and Information Set (HEDIS). In 2013, HEDIS measures covered more than 171 million people from 814 HMOs and 353 PPOs. Below is a description of the denominator for this measure.
Below are the number of health plans and the median eligible population for this measure for the three years of performance data presented.

Rate 1: Timeliness of Prenatal Care
Commercial Rate
YEAR | N Plans | Median Denominator Size per plan
2015 | 380 | 484
2014 | 396 | 502
2013 | 411 | 486
Medicaid Rate
YEAR | N Plans | Median Denominator Size per plan
2015 | 220 | 395
2014 | 217 | 400
2013 | 193 | 436

Rate 2: Postpartum Care
Commercial Rate
YEAR | N Plans | Median Denominator Size per plan
2015 | 377 | 407
2014 | 387 | 503
2013 | 395 | 572
Medicaid Rate
YEAR | N Plans | Median Denominator Size per plan
2015 | 220 | 395
2014 | 217 | 399
2013 | 193 | 436

1b.3. **If no or limited performance data on the measure as specified is reported in 1b2, then provide a summary of data from the literature that indicates opportunity for improvement or overall less than optimal performance on the specific focus of measurement.**
N/A

1b.4. **Provide disparities data from the measure as specified (current and over time) by population group, e.g., by race/ethnicity, gender, age, insurance status, socioeconomic status, and/or disability.** *(This is required for endorsement maintenance. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities include.) This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*
HEDIS data are stratified by type of insurance (e.g. Commercial, Medicaid, Medicare). NCQA does not currently collect performance data stratified by race, ethnicity, or language. Escarce et al. have described in detail the difficulty of collecting valid data on race, ethnicity and language at the health plan level (Escarce, 2011). While not specified in the measure, this measure can also be stratified by demographic variables, such as race/ethnicity or socioeconomic status, in

order to assess the presence of health care disparities. The HEDIS Health Plan Measure Set contains two measures that can assist with stratification to assess health care disparities. The Race/Ethnicity Diversity of Membership and the Language Diversity of Membership measures were designed to promote standardized methods for collecting these data and follow Office of Management and Budget and Institute of Medicine guidelines for collecting and categorizing race/ethnicity and language data. In addition, NCQA's Multicultural Health Care Distinction Program outlines standards for collecting, storing and using race/ethnicity and language data to assess health care disparities. Based on extensive work by NCQA to understand how to promote culturally and linguistically appropriate services among plans and providers, we have many examples of how health plans have used HEDIS measures to design quality improvement programs to decrease disparities in care.

1b.5. **If no or limited data on disparities from the measure as specified is reported in 1b4, then provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement. Include citations.**
Researchers have explored disparities in prenatal and postpartum care and low birth rates, a potential consequence of lack of perinatal care. A 2009 report on disparities in birth outcomes found higher rates of low birth weight amongst non-Hispanic Black women with a low birth rate of 13.6 percent. The low birth rate for non-Hispanic White women was 6.9 percent while Asian/Pacific Islanders were ranked second highest with a low birth rate of 8.3 percent (HHS 2011). Low birth weight infants will experience greater health and social challenges over the course of their lifetime (Goldberg et al. 2007).

CDC data for 2011 show that preterm birth for the Black population was at 16.8 percent, compared to 10.5 percent for non-Hispanic Whites and 11.7 percent for the Hispanic population. The disparity is more evident for early preterm births (infants born before 32 weeks gestation). The Black population has twice as many early preterm births compared to non-Hispanic White and Hispanic populations (CDC 2013). Low-income women are often at risk of early starts to prenatal care due to late enrollment in a health plan during their pregnancy. Delayed enrollment and care initiation results in limited access to prenatal screenings and other resources available within the Medicaid program for the duration of the pregnancy (Egerter S et al. 2002).

Utilization of postpartum care varies based on the level of education attained by the mother. Women with less education are least likely to receive a postpartum care visit. Data from Health and Human Services (HHS) for 2009-2010 show the number of women with less than 12 years of education who received a postpartum care visit was 78.6 percent. Women with 16 or more years of education received postpartum care at a rate of 95.1 percent (HHS 2013b).

1c. High Priority (previously referred to as High Impact)

The measure addresses:

- a specific national health goal/priority identified by DHHS or the National Priorities Partnership convened by NQF;
  OR
- a demonstrated high-priority (high-impact) aspect of healthcare (e.g., affects large numbers of patients and/or has a substantial impact for a smaller population; leading cause of morbidity/mortality; high resource use (current and/or future); severity of illness; and severity of patient/societal consequences of poor quality).

1c.1. **Demonstrated high priority aspect of healthcare**
Affects large numbers, High resource use, Patient/societal consequences of poor quality
1c.2. **If Other:**

1c.3. **Provide epidemiologic or resource use data that demonstrates the measure addresses a high priority aspect of healthcare. List citations in 1c.4.**
Research indicates that early, comprehensive prenatal care can promote healthier pregnancies and reduce the risk of some costly, adverse birth outcomes. Prenatal care allows providers to address key risk factors associated with low birth as smoking, alcohol use, proper weight gain and nutrition and management of existing chronic diseases (CDC 2012). Lack of perinatal care can lead to complications such as preterm birth, which has both negative health consequences and

financial impact on the health care system and patients.

Data from the Centers for Disease Control and Prevention (CDC) indicate 1 in 8 infants are born preterm each year in the United States. These preterm births represented 35 percent of total infant deaths during 2009 (CDC 2013). Children born prematurely are more likely to suffer a variety of motor, cognitive, academic, and behavioral problems, as well as physical health problems in adulthood, such as increased rates of coronary heart disease, stroke, hypertension, and non-insulin dependent diabetes (PHC4 2003; CDC 2013). Preterm infants can experience feeding difficulties resulting from under-developed gastrointestinal systems. Low birth weight babies may face serious health problems upon birth but also are at increased risk of long-term disabilities (March of Dimes 2013).

Premature babies are a driver of excessive maternity costs. Research indicates that early, comprehensive prenatal care can promote healthier pregnancies and reduce the risk of costly, adverse birth outcomes (National Business Group on Health, 2009). A 2007 report by the Institute of Medicine (IOM) indicates preterm birth costs the U.S. $26.2 billion each year. These costs are primarily related to care for the infant but also encompass $1.9 billion in care provided during labor and delivery for the mother (March of Dimes 2013). Moreover, pregnancy complications before delivery account for more than two million hospital days of care each year (CDC 2008).
High costs associated with poor birth outcomes extend beyond the early years of life and into adulthood. Expenditures related to preterm birth have been connected with programs and education services to support preterm infants with developmental delays. A 2007 report from the IOM valued the cost of lost work and pay for persons born prematurely to be $5.6 billion. This amount represents the long-term effects of preterm birth on a person's education and work capabilities (March of Dimes 2013).

The postpartum period is an important time to address the current and future health needs of the mother. During postpartum care visits, it is important to address birth spacing and contraception use to increase the amount of time between pregnancies. Short intervals between pregnancies can cause preterm birth, low birth weight or size small gestational for the infant. In women, there is the increased chance of uterine rupture for women attempting vaginal birth after a cesarean or placental rupture.

The IOM estimates that for every dollar spent on prenatal care, $3.38 is saved in medical costs for low birth weight babies (PHC4 2003). Additionally, there is a large cost savings associated with ensuring adequate prenatal and postpartum care. Every dollar of prenatal care results in expected savings of $3.33 for postnatal care and $4.63 in long-term morbidity costs (Lu M et al. 2000). Adequate and timely prenatal care not only improves the likelihood of a healthy birth but also provides a higher probability of a prosperous future for the infant. The societal costs of prematurity add greater incentive to ensure women are initiating and receiving adequate prenatal care.

1c.4. Citations for data demonstrating high priority provided in 1a.3

1c.5. If a PRO-PM (e.g. HRQoL/functional status, symptom/burden, experience with care, health-related behaviors), provide evidence that the target population values the measured PRO and finds it meaningful. *(Describe how and from whom their input was obtained.)*

2. Reliability and Validity—Scientific Acceptability of Measure Properties

Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. ***Measures must be judged to meet the subcriteria for both reliability and validity to pass this criterion and be evaluated against the remaining criteria.***

2a.1. Specifications The measure is well defined and precisely specified so it can be implemented consistently within and across organizations and allows for comparability. eMeasures should be specified in the Health Quality Measures Format (HQMF) and the Quality Data Model (QDM).

De.5. **Subject/Topic Area** *(check all the areas that apply):*
Perinatal and Reproductive Health, Perinatal and Reproductive Health : Perinatal, Perinatal and Reproductive Health : Screening

De.6. **Cross Cutting Areas** *(check all the areas that apply):*
Access

S.1. **Measure-specific Web Page** *(Provide a URL link to a web page specific for this measure that contains current detailed specifications including code lists, risk model details, and supplemental materials. Do not enter a URL linking to a home page or to general information.)*
N/A

S.2a. **If this is an eMeasure**, HQMF specifications must be attached. Attach the zipped output from the eMeasure authoring tool (MAT) - if the MAT was not used, contact staff. (Use the specification fields in this online form for the plain-language description of the specifications)
This is not an eMeasure **Attachment:**

S.2b. **Data Dictionary, Code Table, or Value Sets** *(and risk model codes and coefficients when applicable) must be attached. (Excel or csv file in the suggested format preferred - if not, contact staff)*
Attachment **Attachment:** 1517_PPC_Value_Sets.xlsx

S.3. **For endorsement maintenance**, please briefly describe any changes to the measure specifications since last endorsement date and explain the reasons.
SUMMARY OF CHANGES
• Deleted the use of infant claims to identify deliveries.
• Clarified the tests that must be included to meet criteria for an obstetric panel in the medical record specification. These are as follows: hematocrit, differential WBC count, platelet count, hepatitis B surface antigen, rubella antibody, syphilis test, RBC antibody screen, Rh and ABO blood typing.

S.4. **Numerator Statement** *(Brief, narrative description of the measure focus or what is being measured about the target population, i.e., cases from the target population with the target process, condition, event, or outcome)*
IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
This measure assesses whether pregnant women had timely prenatal and postpartum care visits. It has two rates, one assessing the timeliness of prenatal visits, and one assessing the timeliness of postpartum visits.

S.5. **Time Period for Data** *(What is the time period in which data will be aggregated for the measure, e.g., 12 mo, 3 years, look back to August for flu vaccination? Note if there are different time periods for the numerator and denominator.)*
12 months

S.6. **Numerator Details** *(All information required to identify and calculate the cases from the target population with the target process, condition, event, or outcome such as definitions, specific data collection items/responses, code/value sets*

– Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)
IF an OUTCOME MEASURE, describe how the observed outcome is identified/counted. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
Administrative Specifications

Timeliness of Prenatal Care
A prenatal visit in the first trimester or within 42 days of enrollment, depending on the date of enrollment in the organization and the gaps in enrollment during the pregnancy. Include only visits that occur while the member was enrolled.
Follow the steps below to identify the numerator.
Step 1: Determine enrollment status during the first trimester. For all women in the eligible population, identify those who were enrolled on or before 280 days prior to delivery (or estimated date of delivery [EDD]). For these women, proceed to step 2.
For women not enrolled on or before 280 days prior to delivery (or EDD), who were therefore pregnant at the time of enrollment, proceed to step 3.
Step 2: Determine continuous enrollment for the first trimester. Identify women from step 1 who were continuously enrolled during the first trimester (176–280 days prior to delivery [or EDD]), with no gaps in enrollment. For these women, determine numerator compliance using the decision rules for Identifying Prenatal Care For Women Continuously Enrolled During the First Trimester.
For women who were not continuously enrolled during the first trimester (e.g., had a gap between 176 and 280 days before delivery), proceed to step 3.
Step 3: Determine the start date of the last enrollment segment (i.e., the enrollment segment during the pregnancy with the start date that is closest to the delivery date).
For women whose last enrollment started on or between 219 and 279 days before delivery, proceed to step 4.
For women whose last enrollment started less than 219 days before delivery, proceed to step 5.
Step 4: Determine numerator compliance. If the last enrollment segment started on or between 219 and 279 days before delivery, determine numerator compliance using the instructions for Identifying Prenatal Care for Women Not Continuously Enrolled During the First Trimester and find a visit between the last enrollment start date and 176 days before delivery.
Step 5: Determine numerator compliance. If the last enrollment segment started less than 219 days before delivery (i.e., between 219 days before delivery and the day of delivery), determine numerator compliance using the instructions for Identifying Prenatal Care for Women Not Continuously Enrolled During the First Trimester and find a visit within 42 days after enrollment.

Identifying Prenatal Care for Women Continuously Enrolled During the First Trimester
Decision Rule 1
Either of the following during the first trimester, where the practitioner type is an OB/GYN or other prenatal care practitioner or PCP meets criteria:
• A bundled service (Prenatal Bundled Services Value Set) where the organization can identify the date when prenatal care was initiated (because bundled service codes are used on the date of delivery, these codes may be used only if the claim form indicates when prenatal care was initiated).
• A visit for prenatal care (Stand Alone Prenatal Visits Value Set).
Decision Rule 2
Any of the following during the first trimester, where the practitioner type for the prenatal visit is an OB/GYN or other prenatal care practitioner, meet criteria:
• A prenatal visit (Prenatal Visits Value Set) with an obstetric panel (Obstetric Panel Value Set).
• A prenatal visit (Prenatal Visits Value Set) with an ultrasound (echocardiography) of the pregnant uterus (Prenatal Ultrasound Value Set).
• A prenatal visit (Prenatal Visits Value Set) with a pregnancy-related diagnosis code (Pregnancy Diagnosis Value Set).
• A prenatal visit (Prenatal Visits Value Set) with all of the following:
– Toxoplasma (Toxoplasma Antibody Value Set).

– Rubella (Rubella Antibody Value Set).
– Cytomegalovirus (Cytomegalovirus Antibody Value Set).
– Herpes simplex (Herpes Simplex Antibody Value Set).
• A prenatal visit (Prenatal Visits Value Set) with rubella (Rubella Antibody Value Set) and ABO (ABO Value Set).
• A prenatal visit (Prenatal Visits Value Set) with rubella (Rubella Antibody Value Set) and Rh (Rh Value Set).
• A prenatal visit (Prenatal Visits Value Set) with rubella (Rubella Antibody Value Set) and ABO/Rh (ABO and Rh Value Set).

Decision Rule 3

Any of the following during the first trimester, where the practitioner type is a PCP, meet criteria:
• A prenatal visit (Prenatal Visits Value Set) with a pregnancy-related diagnosis code (Pregnancy Diagnosis Value Set) and an obstetric panel (Obstetric Panel Value Set).
• A prenatal visit (Prenatal Visits Value Set) with a pregnancy-related diagnosis code (Pregnancy Diagnosis Value Set) and an ultrasound (echocardiography) of the pregnant uterus (Prenatal Ultrasound Value Set).
• A prenatal visit (Prenatal Visits Value Set) with a pregnancy-related diagnosis code (Pregnancy Diagnosis Value Set) and all of the following:
– Toxoplasma (Toxoplasma Antibody Value Set).
– Rubella (Rubella Antibody Value Set).
– Cytomegalovirus (Cytomegalovirus Antibody Value Set).
– Herpes simplex (Herpes Simplex Antibody Value Set).
• A prenatal visit (Prenatal Visits Value Set) with a pregnancy-related diagnosis code (Pregnancy Diagnosis Value Set) and rubella (Rubella Antibody Value Set) and ABO (ABO Value Set).
• A prenatal visit (Prenatal Visits Value Set) with a pregnancy-related diagnosis code (Pregnancy Diagnosis Value Set) and rubella (Rubella Antibody Value Set) and Rh (Rh Value Set).
• A prenatal visit (Prenatal Visits Value Set) with a pregnancy-related diagnosis code (Pregnancy Diagnosis Value Set) and rubella (Rubella Antibody Value Set) and ABO/Rh (ABO and Rh Value Set).
• A prenatal visit (Prenatal Visits Value Set) with any internal organization code for LMP or EDD with an obstetrical history.
• A prenatal visit (Prenatal Visits Value Set) with any internal organization code for LMP or EDD with risk assessment and counseling/education.

Note: For Decision Rule 3 criteria that require a prenatal visit code (Prenatal Visits Value Set) and a pregnancy-related diagnosis code (Pregnancy Diagnosis Value Set), codes must be on the same claim.

Identifying Prenatal Care for Women Not Continuously Enrolled During the First Trimester

Any of the following, where the practitioner type is an OB/GYN or other prenatal care practitioner or PCP, meet criteria:
• A bundled service (Prenatal Bundled Services Value Set) where the organization can identify the date when prenatal care was initiated (because bundled service codes are used on the date of delivery, these codes may be used only if the claim form indicates when prenatal care was initiated).
• A visit for prenatal care (Stand Alone Prenatal Visits Value Set).
• A prenatal visit (Prenatal Visits Value Set) with an ultrasound (echocardiography) of the pregnant uterus (Prenatal Ultrasound Value Set).
• A prenatal visit (Prenatal Visits Value Set) with a principal pregnancy-related diagnosis code (Pregnancy Diagnosis Value Set).

Note: For criteria that require a prenatal visit code (Prenatal Visits Value Set) and a pregnancy-related diagnosis code (Pregnancy Diagnosis Value Set), codes must be on the same claim. Criteria for identifying prenatal care for women who were not continuously enrolled during the first trimester allow more flexibility than criteria for women who were continuously enrolled.

Postpartum Care

A postpartum visit for a pelvic exam or postpartum care on or between 21 and 56 days after delivery. Any of the following meet criteria:
• A postpartum visit (Postpartum Visits Value Set).
• Cervical cytology (Cervical Cytology Value Set).

• A bundled service (Postpartum Bundled Services Value Set) where the organization can identify the date when postpartum care was rendered (because bundled service codes are used on the date of delivery, not on the date of the postpartum visit, these codes may be used only if the claim form indicates when postpartum care was rendered).
Note: The practitioner requirement only applies to the Hybrid Specification. The organization is not required to identify practitioner type in administrative data.

Medical Record Specification

Timeliness of Prenatal Care
A prenatal visit in the first trimester or within 42 days of enrollment, depending on the date of enrollment in the organization and gaps in enrollment during the pregnancy. Include only visits that occurred while the member was enrolled.
Prenatal care visit to an OB/GYN or other prenatal care practitioner or PCP. For visits to a PCP, a diagnosis of pregnancy must be present. Documentation in the medical record must include a note indicating the date when the prenatal care visit occurred, and evidence of one of the following.
• A basic physical obstetrical examination that includes auscultation for fetal heart tone, or pelvic exam with obstetric observations, or measurement of fundus height (a standardized prenatal flow sheet may be used).
• Evidence that a prenatal care procedure was performed, such as:
– Screening test in the form of an obstetric panel (must include all of the following: hematocrit, differential WBC count, platelet count, hepatitis B surface antigen, rubella antibody, syphilis test, RBC antibody screen, Rh and ABO blood typing), or
– TORCH antibody panel alone, or
– A rubella antibody test/titer with an Rh incompatibility (ABO/Rh) blood typing, or
– Echography of a pregnant uterus.
• Documentation of LMP or EDD in conjunction with either of the following.
– Prenatal risk assessment and counseling/education.
– Complete obstetrical history.
Note: For women whose last enrollment segment was after 219 days prior to delivery (i.e., between 219 days prior to delivery and the day of delivery) and women who had a gap during the first trimester, count documentation of a visit to an OB/GYN, family practitioner or other PCP with a principal diagnosis of pregnancy.

Postpartum Care
A postpartum visit for a pelvic exam or postpartum care on or between 21 and 56 days after delivery, as documented through either administrative data or medical record review.
Postpartum visit to an OB/GYN practitioner or midwife, family practitioner or other PCP on or between 21 and 56 days after delivery. Documentation in the medical record must include a note indicating the date when a postpartum visit occurred and one of the following.
• Pelvic exam.
• Evaluation of weight, BP, breasts and abdomen.
– Notation of “breastfeeding” is acceptable for the “evaluation of breasts” component.
• Notation of postpartum care, including, but not limited to:
– Notation of “postpartum care,” “PP care,” “PP check,” “6-week check.”
– A preprinted “Postpartum Care” form in which information was documented during the visit.

For both rates:
• Services that occur over multiple visits count toward this measure if all services are within the time frame established in the measure. Ultrasound and lab results alone are not considered a visit; they must be linked to an office visit with an appropriate practitioner in order to count for this measure.
• NCQA defines a PCP and OB/GYN and other prenatal practitioners as including:
• Physicians certified as obstetricians or gynecologists by the American Medical Specialties Board of Obstetrics or Gynecology or the American Osteopathic Association; or, if not certified, who successfully completed an accredited program of graduate medical or osteopathic education in obstetrics and gynecology.

• Certified nurse midwives and nurse practitioners who deliver prenatal care services in a specialty setting (under the direction of an OB/GYN certified or accredited provider).

S.7. **Denominator Statement** *(Brief, narrative description of the target population being measured)*
The percentage of deliveries of live births between November 6 of the year prior to the measurement year and November 5 of the measurement year.

S.8. **Target Population Category** *(Check all the populations for which the measure is specified and tested if any):*
Maternal Health

S.9. **Denominator Details** *(All information required to identify and calculate the target population/denominator such as definitions, specific data collection items/responses , code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
Product Lines: Commercial, Medicaid (report each product line separately).
Continuous enrollment: 43 days prior to delivery through 56 days after delivery.
Allowable gap: No allowable gap during the continuous enrollment period.
Anchor date: Date of delivery.
Benefit: Medical.
Event/ diagnosis: Delivered a live birth on or between November 6 of the year prior to the measurement year and November 5 of the measurement year. Include women who delivered in any setting.
Multiple births. Women who had two separate deliveries (different dates of service) between November 6 of the year prior to the measurement year and November 5 of the measurement year count twice. Women who had multiple live births during one pregnancy count once.

Follow the steps below to identify the eligible population, which is the denominator for both rates.
Step 1: Identify deliveries. Identify all women with a delivery (Deliveries Value Set) between November 6 of the year prior to the measurement year and November 5 of the measurement year.
Step 2: Exclude non-live births (Non-live Births Value Set).
Step 3: Identify continuous enrollment. Determine if enrollment was continuous between 43 days prior to delivery and 56 days after delivery, with no gaps.

S.10. **Denominator Exclusions** *(Brief narrative description of exclusions from the target population)*
Non-live births

S.11. **Denominator Exclusion Details** *(All information required to identify and calculate exclusions from the denominator such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
See corresponding Excel document for the Non-live Births Value Set.

S.12. **Stratification Details/Variables** *(All information required to stratify the measure results including the stratification variables, definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format with at S.2b)*
N/A

S.13. **Risk Adjustment Type** (Select type. Provide specifications for risk stratification in S.12 and for statistical model in S.14-15)
No risk adjustment or risk stratification
If other:

S.14. **Identify the statistical risk model method and variables** *(Name the statistical method - e.g., logistic regression and list all the risk factor variables. Note - risk model development and testing should be addressed with measure testing under Scientific Acceptability)*

N/A

S.15. **Detailed risk model specifications** *(must be in attached data dictionary/code list Excel or csv file. Also indicate if available at measure-specific URL identified in S.1.)*
Note: Risk model details (including coefficients, equations, codes with descriptors, definitions), should be provided on a separate worksheet in the suggested format in the Excel or csv file with data dictionary/code lists at S.2b.

S.15a. **Detailed risk model specifications** (*if not provided in excel or csv file at S.2b*)
N/A

S.16. **Type of score:**
Rate/proportion
If other:

S.17. **Interpretation of Score** *(Classifies interpretation of score according to whether better quality is associated with a higher score, a lower score, a score falling within a defined interval, or a passing score)*
Better quality = Higher score

S.18. **Calculation Algorithm/Measure Logic** *(Describe the calculation of the measure score as an ordered sequence of steps including identifying the target population; exclusions; cases meeting the target process, condition, event, or outcome; aggregating data; risk adjustment; etc.)*
Step 1: Calculate the eligible population following the instructions in the denominator details listed in section S.9.
Step 2: Remove the exclusions identified in section S.10.
Step 3: Calculate the numerator for Rate 1 following the instructions in the numerator details listed in section S.6.
Step 4: Divide the numerator from Step 3 by the denominator from Step 2 to determine Rate 1.
Step 5: Calculate the numerator for Rate 2 following the instructions in the numerator details listed in section S.6.
Step 6: Divide the numerator from Step 5 by the denominator from Step 2 to determine Rate 2.

S.19. **Calculation Algorithm/Measure Logic Diagram URL or Attachment** *(You also may provide a diagram of the Calculation Algorithm/Measure Logic described above at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
No diagram provided

S.20. **Sampling** *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*
IF a PRO-PM, identify whether (and how) proxy responses are allowed.
N/A

S.21. **Survey/Patient-reported data** *(If measure is based on a survey, provide instructions for conducting the survey and guidance on minimum response rate.)*
IF a PRO-PM, specify calculation of response rates to be reported with performance measure results.
N/A

S.22. **Missing data** (specify how missing data are handled, e.g., imputation, delete case.)
Required for Composites and PRO-PMs.
N/A

S.23. **Data Source** *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*
If other, please describe in S.24.
Administrative claims, Electronic Clinical Data, Paper Medical Records

S.24. **Data Source or Collection Instrument** *(Identify the specific data source/data collection instrument e.g. name of database, clinical registry, collection instrument, etc.)*
IF a PRO-PM, identify the specific PROM(s); and standard methods, modes, and languages of administration.
This measure is based on administrative claims and medical record documentation collected in the course of providing care to health plan members. NCQA collects the Healthcare Effectiveness Data and Information Set (HEDIS) data for this measure directly from Health Management Organizations and Preferred Provider Organizations via NCQA's online data submission system.

S.25. **Data Source or Collection Instrument** *(available at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
No data collection instrument provided

S.26. **Level of Analysis** *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*
Health Plan, Integrated Delivery System

S.27. **Care Setting** *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*
Ambulatory Care : Clinician Office/Clinic
If other:

S.28. **COMPOSITE Performance Measure** - Additional Specifications *(Use this section as needed for aggregation and weighting rules, or calculation of individual performance measures if not individually endorsed.)*
N/A

2a. Reliability – See attached Measure Testing Submission Form
2b. Validity – See attached Measure Testing Submission Form
1517_MeasureTesting_MSF5.0_Data-635278463206876303-635920201840831482.doc

2. RELIABILITY & VALIDITY - SCIENTIFIC ACCEPTABILITY OF MEASURE PROPERTIES

Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. **(evaluation criteria)**

Measure testing must demonstrate adequate reliability and validity in order to be recommended for endorsement. Testing may be conducted for data elements and/or the computed measure score. Testing information and results should be entered in the appropriate field. Supplemental materials may be referenced or attached in item 2.1. See guidance on measure testing.

2a2. Reliability Testing. (*Reliability testing was conducted with appropriate method, scope, and adequate demonstration of reliability*.)

2a2.1 Data/Sample *(Description of the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)*:

The reliability metric for each measure was calculated separately for Commercial and Medicaid plans where applicable using 2010 data.

2a2.2 Analytic Method *(Describe method of reliability testing & rationale)*:

Reliability was estimated by using the beta-binomial model. Beta-binomial is a better fit when estimating the reliability of simple pass/fail rate measures as is the case with most HEDIS® health plan measures. The beta-binomial model assumes the plan score is a binomial random variable conditional on the plan´s true value that comes from the beta distribution. The beta distribution is usually defined by two parameters, alpha and beta. Alpha and beta can be thought of as intermediate calculations to get to the needed variance estimates. The beta distribution can be symmetric, skewed or even U-shaped.

Reliability used here is the ratio of signal to noise. The signal in this case is the proportion of the variability in measured performance that can be explained by real differences in performance.

2a2.3 Testing Results *(Reliability statistics, assessment of adequacy in the context of norms for the test conducted)*:

Rate - Timeliness of Prenatal Care:

Reliability - Commercial Plans: 0.9961

Reliability - Medicaid Plans: 0.9564

Rate - Postpartum Care:

Reliability - Commercial Plans: 0.9944

Reliability - Medicaid Plans: 0.9217

A reliability of zero implies that all the variability in a measure is attributable to measurement error. A reliability of one implies that all the variability is attributable to real differences in performance. The higher the reliability score, the greater is the confidence with which one can distinguish the performance of one plan from another. A reliability score greater than or equal to 0.7 is considered very good.

2b. VALIDITY. Validity, Testing, including all Threats to Validity: H☐ M☐ L☐ I ☐

2b1.1 **Describe how the measure specifications** *(measure focus, target population, and exclusions)* **are consistent with the evidence cited in support of the measure focus (***criterion 1c)* **and identify any differences from the evidence:**

2b2. Validity Testing. (*Validity testing was conducted with appropriate method, scope, and adequate demonstration of validity.)*

2b2.1 **Data/Sample** *(Description of the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)***:**

expert panel

2b2.2 **Analytic Method** *(Describe method of validity testing and rationale; if face validity, describe systematic assessment)***:**

NCQA tested the measure for face validity using a panel of stakeholders with specific expertise in measurement and child health care. This panel included representatives from key stakeholder groups, including pediatricians, family physicians, health plans, state Medicaid agencies and researchers. Experts reviewed the results of the field test and assessed whether the results were consistent with expectations, whether the measure represented quality care, and whether we were measuring the most important aspect of care in this area.

2b2.3 **Testing Results** *(Statistical results, assessment of adequacy in the context of norms for the test conducted; if face validity, describe results of systematic assessment)***:**

This measure was deemed valid by the expert panel.

POTENTIAL THREATS TO VALIDITY. (*All potential threats to validity were appropriately tested with adequate results.*)

2b3. Measure Exclusions. (*Exclusions were supported by the clinical evidence in 1c or appropriately tested with results demonstrating the need to specify them.*)

2b3.1 **Data/Sample for analysis of exclusions** *(Description of the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)***:**

NA

2b3.2 **Analytic Method** (*Describe type of analysis and rationale for examining exclusions, including exclusion related to patient preference)***:**

NA

2b3.3 **Results** *(Provide statistical results for analysis of exclusions, e.g., frequency, variability, sensitivity analyses):*

NA

2b4. Risk Adjustment Strategy. (*For outcome measures, adjustment for differences in case mix (severity) across measured entities was appropriately tested with adequate results.*)

2b4.1 **Data/Sample** *(Description of the data or sample including number of measured entities; number of patients; dates of*

data; if a sample, characteristics of the entities included):

NA

2b4.2 **Analytic Method (***Describe methods and rationale for development and testing of risk model or risk stratification including selection of factors/variables)*:

NA

2b4.3 **Testing Results** *(Statistical risk model: Provide quantitative assessment of relative contribution of model risk factors; risk model performance metrics including cross-validation discrimination and calibration statistics, calibration curve and risk decile plot, and assessment of adequacy in the context of norms for risk models. Risk stratification: Provide quantitative assessment of relationship of risk factors to the outcome and differences in outcomes among the strata)*:

NA

2b4.4 **If outcome or resource use measure is not risk adjusted, provide rationale and analyses to justify lack of adjustment:** The measure assesses utilization and access in a general population; risk adjustment is not indicated.

2b5. Identification of Meaningful Differences in Performance. (*The performance measure scores were appropriately analyzed and discriminated meaningful differences in quality.*)

2b5.1 **Data/Sample** *(Describe the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)*:

HEDIS National Data

2b5.2 **Analytic Method** *(Describe methods and rationale to identify statistically significant and practically/meaningfully differences in performance)*:

Comparison of means and percentiles; analysis of variance against established benchmarks; if sample size is >400, we would use an analysis of variance

2b5.3 **Results** *(Provide measure performance results/scores, e.g., distribution by quartile, mean, median, SD, etc.; identification of statistically significant and meaningfully differences in performance)*:

Frequency of Ongoing Prenatal Care

National Means

HEDIS 2006 Data

<21%

13.52

HEDIS 2007 Data

12.36

>= 81%

HEDIS 2006 Data

58.6

HEDIS 2007 Data

59.59

21-40%

HEDIS 2006 Data

6.04

HEDIS 2007 Data

6.63

41-60%

HEDIS 2006 Data

7.84

HEDIS 2007 Data

7.74

61-80%

HEDIS 2006 Data

14.1

HEDIS 2007 Data

13.85

Prenatal and Postpartum Care

Postpartum Care

HEDIS 2006

59.08

HEDIS 2007

58.6

Timeliness of Prenatal Care

HEDIS 2006

81.24

HEDIS 2007

81.37

2b6. Comparability of Multiple Data Sources/Methods. (*If specified for more than one data source, the various approaches result in comparable scores.*)

2b6.1 Data/Sample *(Describe the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)***:**

2b6.2 Analytic Method *(Describe methods and rationale for testing comparability of scores produced by the different data sources specified in the measure)***:**

2b6.3 Testing Results *(Provide statistical results, e.g., correlation statistics, comparison of rankings; assessment of adequacy in the context of norms for the test conducted)***:**

2c. Disparities in Care: H☐ M☐ L☐ I☐ NA☐ (*If applicable, the measure specifications allow identification of disparities.*)

2c.1 If measure is stratified for disparities, provide stratified results *(Scores by stratified categories/cohorts)***:** The measure is not stratified to detect disparities.

2c.2 If disparities have been reported/identified (e.g., in 1b), but measure is not specified to detect disparities, please explain:

NA

2.1-2.3 Supplemental Testing Methodology Information:

Steering Committee: Overall, was the criterion, *Scientific Acceptability of Measure Properties*, met?

(*Reliability and Validity must be rated moderate or high*) Yes☐ No☐

Provide rationale based on specific subcriteria:

If the Committee votes No, STOP

3. Feasibility

Extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

3a. Byproduct of Care Processes

For clinical measures, the required data elements are routinely generated and used during care delivery (e.g., blood pressure, lab test, diagnosis, medication order).

3a.1. **Data Elements Generated as Byproduct of Care Processes.**

Generated or collected by and used by healthcare personnel during the provision of care (e.g., blood pressure, lab value, diagnosis, depression score), Coded by someone other than person obtaining original information (e.g., DRG, ICD-9 codes on claims), Abstracted from a record by someone other than person obtaining original information (e.g., chart abstraction for quality measure or registry)

If other:

3b. Electronic Sources

The required data elements are available in electronic health records or other electronic sources. If the required data are not in electronic health records or existing electronic sources, a credible, near-term path to electronic collection is specified.

3b.1. **To what extent are the specified data elements available electronically in defined fields?** *(i.e., data elements that are needed to compute the performance measure score are in defined, computer-readable fields)*

Some data elements are in defined fields in electronic sources

3b.2. **If ALL the data elements needed to compute the performance measure score are not from electronic sources, specify a credible, near-term path to electronic capture, OR provide a rationale for using other than electronic sources.**

To allow for widespread reporting across health plans and health care practices, this measure is collected through multiple data sources (administrative data, electronic clinical data, paper records). We anticipate as electronic health records become more widespread the reliance on paper record review will decrease.

3b.3. **If this is an eMeasure, provide a summary of the feasibility assessment in an attached file or make available at a measure-specific URL.**

Attachment:

3c. Data Collection Strategy

Demonstration that the data collection strategy (e.g., source, timing, frequency, sampling, patient confidentiality, costs associated with fees/licensing of proprietary measures) can be implemented (e.g., already in operational use, or testing demonstrates that it is ready to put into operational use). For eMeasures, a feasibility assessment addresses the data elements and measure logic and demonstrates the eMeasure can be implemented or feasibility concerns can be adequately addressed.

3c.1. **Describe what you have learned/modified as a result of testing and/or operational use of the measure regarding data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, time and cost of data collection, other feasibility/implementation issues.**

IF a PRO-PM, consider implications for both individuals providing PROM data (patients, service recipients, respondents) and those whose performance is being measured.

NCQA recognizes that, despite the clear specifications defined for HEDIS measures, data collection and calculation methods may vary, and other errors may affect the results. Thus, NCQA conducts an independent audit of all HEDIS collection and reporting processes, as well as an audit of the data which are manipulated by those processes, in order to verify that HEDIS specifications are met. NCQA has developed a precise, standardized methodology for verifying the integrity of HEDIS collection and calculation processes through a two-part program consisting of an overall information systems capabilities assessment followed by an evaluation of the MCO´s ability to comply with HEDIS specifications.

NCQA-certified auditors using standard audit methodologies will help enable purchasers to make more reliable "apples-to-apples" comparisons between health plans.

The HEDIS Compliance Audit addresses the following functions:
1) information practices and control procedures
2) sampling methods and procedures
3) data integrity
4) compliance with HEDIS specifications
5) analytic file production
6) reporting and documentation

In addition to the HEDIS Audit, NCQA provides a system to allow “real-time” feedback from measure users. Our Policy Clarification Support System receives thousands of inquiries each year on over 100 measures. Through this system NCQA responds immediately to questions and identifies possible errors or inconsistencies in the implementation of the measure. This system is vital to the regular re-evaluation of NCQA measures.
Input from NCQA auditing and the Policy Clarification Support System informs the annual updating of all HEDIS measures including updating value sets and clarifying the specifications. Measures are re-evaluated on a periodic basis and when there is a significant change in evidence. During re-evaluation information from NCQA auditing and Policy Clarification Support System is used to inform evaluation of the scientific soundness and feasibility of the measure.

3c.2. Describe any fees, licensing, or other requirements to use any aspect of the measure as specified *(e.g., value/code set, risk model, programming code, algorithm)*.
Broad public use and dissemination of these measures is encouraged and NCQA has agreed with NQF that noncommercial uses do not require the consent of the measure developer. Use by health care physicians in connection with their own practices is not commercial use. Commercial use of a measure requires the prior written consent of NCQA. As used herein, “commercial use” refers to any sale, license or distribution of a measure for commercial gain, or incorporation of a measure into any product or service that is sold, licensed or distributed for commercial gain, even if there is no actual charge for inclusion of the measure.

4. Usability and Use

Extent to which potential audiences (e.g., consumers, purchasers, providers, policy makers) are using or could use performance results for both accountability and performance improvement to achieve the goal of high-quality, efficient healthcare for individuals or populations.

4a. Accountability and Transparency

Performance results are used in at least one accountability application within three years after initial endorsement and are publicly reported within six years after initial endorsement (or the data on performance results are available). If not in use at the time of initial endorsement, then a credible plan for implementation within the specified timeframes is provided.

4.1. Current and Planned Use

NQF-endorsed measures are expected to be used in at least one accountability application within 3 years and publicly reported within 6 years of initial endorsement in addition to performance improvement.

Planned	Current Use (for current use provide URL)
	Public Reporting NCQA Quality Compass http://www.ncqa.org/tabid/177/Default.aspx NCQA State of Health Care Quality Medicaid Child Core Set http://www.ncqa.org/tabid/836/Default.aspx Quality Improvement with Benchmarking (external benchmarking to multiple organizations) NCQA Quality Compass NCQA State of Health Care Quality Medicaid Child Core Set http://www.ncqa.org/tabid/177/Default.aspx

4a.1. For each CURRENT use, checked above, provide:

- Name of program and sponsor
- Purpose
- Geographic area and number and percentage of accountable entities and patients included

QUALITY COMPASS: This measure is used in Quality Compass which is an indispensable tool used for selecting a health plan, conducting competitor analysis, examining quality improvement and benchmarking plan performance. Provided in this tool is the ability to generate custom reports by selecting plans, measures, and benchmarks (averages and percentiles) for up to three trended years. Results in table and graph formats offer simple comparison of plans' performance against competitors or benchmarks.

STATE OF HEALTH CARE QUALITY REPORT: This measure is publically reported nationally and by geographic regions in the NCQA State of Health Care annual report. This annual report published by NCQA summarizes findings on quality of care. In 2012 the report included measures on 11.5 million Medicare Advantage beneficiaries in 455 Medicare Advantage health plans, 99.4 million members in 404 commercial health plans, and 14.3 million Medicaid beneficiaries in 136 plans across 50 states.

MEDICAID CHILD CORE SET: These are a core set of health quality measures for Medicaid-enrolled children. The Medicaid Child Core Set was identified by the Centers for Medicare & Medicaid (CMS) in partnership with the Agency for HealthCare Research and Quality (AHRQ). The data collected from these measures will help CMS to better understand the quality of health care that children enrolled in Medicaid receive nationally. On December 29, 2009, the Secretary posted for public comment in the Federal Register, an initial core set of 24 children´s health care quality measures for voluntary use by Medicaid and CHIP programs. The CHIPRA legislation provides that the Secretary shall issue updates to

the Child Core Set beginning in January 2013 and annually thereafter. CMS worked with the National Quality Forum's (NQF) Measures Application Partnership (MAP) to review the Child Core Set and to identify ways to improve it. State data derived from the core measures are part of the Secretary´s Annual Report on the Quality of Care for Children in Medicaid and CHIP.

4a.2. **If not currently publicly reported OR used in at least one other accountability application (e.g., payment program, certification, licensing) what are the reasons?** (*e.g., Do policies or actions of the developer/steward or accountable entities restrict access to performance results or impede implementation?*)
N/A

4a.3. **If not currently publicly reported OR used in at least one other accountability application, provide a credible plan for implementation within the expected timeframes -- any accountability application within 3 years and publicly reported within 6 years of initial endorsement.** (*Credible plan includes the specific program, purpose, intended audience, and timeline for implementing the measure within the specified timeframes. A plan for accountability applications addresses mechanisms for data aggregation and reporting.*)
N/A

4b. Improvement

Progress toward achieving the goal of high-quality, efficient healthcare for individuals or populations is demonstrated. If not in use for performance improvement at the time of initial endorsement, then a credible rationale describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.

4b.1. **Progress on Improvement. (Not required for initial endorsement unless available.)**
Performance results on this measure (current and over time) should be provided in 1b.2 and 1b.4. Discuss:

- **Progress (trends in performance results, number and percentage of people receiving high-quality healthcare)**
- **Geographic area and number and percentage of accountable entities and patients included**

Performance results (listed in 1b.2 and 1b.4) show that rates have been steady over the past three years among commercial and Medicaid plans. However, there continues to be variation in the Prenatal and Postpartum Care measure by each of its two rates across health plans. In 2015, for the Prenatal Care rate, there was a 28 percentage point difference between plans in the 10th percentile vs the 90th percentile for commercial plans and 23 percentage point difference for Medicaid plans. For the Postpartum Care rate, there was a 42 percentage point difference between plans in the 10th percentile vs the 90th percentile for commercial plans and 23 percentage point difference for Medicaid plans. These gaps in performance underscore the opportunity for improvement.

4b.2. **If no improvement was demonstrated, what are the reasons? If not in use for performance improvement at the time of initial endorsement, provide a credible rationale that describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.**
It is not clear why improvement has not occurred, though as noted there continues to be variation among health care plans. The measure is actively used in programs for both health plan and state reporting. The measure assesses pregnant women's access to care; it is possible that health insurance expansion will increase access to prenatal and postpartum care over time.

4c. Unintended Consequences

The benefits of the performance measure in facilitating progress toward achieving high-quality, efficient healthcare for individuals or populations outweigh evidence of unintended negative consequences to individuals or populations (if such evidence exists).

4c.1. **Were any unintended negative consequences to individuals or populations identified during testing; OR has evidence of unintended negative consequences to individuals or populations been reported since implementation? If so, identify the negative unintended consequences and describe how benefits outweigh them or actions taken to mitigate them.**
There were no identified unintended consequences for this measure during testing or since implementation.

5. Comparison to Related or Competing Measures

If a measure meets the above criteria and there are endorsed or new related measures (either the same measure focus or the same target population) or competing measures (both the same measure focus and the same target population), the measures are compared to address harmonization and/or selection of the best measure.

5. Relation to Other NQF-endorsed Measures
Are there related measures (conceptually, either same measure focus or target population) or competing measures (conceptually both the same measure focus and same target population)? If yes, list the NQF # and title of all related and/or competing measures.
Yes

5.1a. **List of related or competing measures (selected from NQF-endorsed measures)**
1391 : Frequency of Ongoing Prenatal Care (FPC)

5.1b. **If related or competing measures are not NQF endorsed please indicate measure title and steward.**

5a. Harmonization

The measure specifications are harmonized with related measures;
OR
The differences in specifications are justified

5a.1. **If this measure conceptually addresses EITHER the same measure focus OR the same target population as NQF-endorsed measure(s):**
Are the measure specifications completely harmonized?
Yes

5a.2. **If the measure specifications are not completely harmonized, identify the differences, rationale, and impact on interpretability and data collection burden.**

5b. Competing Measures

The measure is superior to competing measures (e.g., is a more valid or efficient way to measure);
OR
Multiple measures are justified.

5b.1. **If this measure conceptually addresses both the same measure focus and the same target population as NQF-endorsed measure(s):**
Describe why this measure is superior to competing measures (e.g., a more valid or efficient way to measure quality); OR provide a rationale for the additive value of endorsing an additional measure. (Provide analyses when possible.)

Appendix

A.1 Supplemental materials may be provided in an appendix. All supplemental materials (such as data collection instrument or methodology reports) should be organized in one file with a table of contents or bookmarks. If material pertains to a specific submission form number, that should be indicated. Requested information should be provided in the submission form and required attachments. There is no guarantee that supplemental materials will be reviewed.

Attachment:

Contact Information

Co.1 **Measure Steward (Intellectual Property Owner):** National Committee for Quality Assurance
Co.2 **Point of Contact:** Bob, Rehm, nqf@ncqa.org, 202-955-1728-
Co.3 **Measure Developer if different from Measure Steward:** National Committee for Quality Assurance
Co.4 **Point of Contact:** Bob, Rehm, nqf@ncqa.org, 202-955-1728-

Additional Information

Ad.1 **Workgroup/Expert Panel involved in measure development**
Provide a list of sponsoring organizations and workgroup/panel members' names and organizations. Describe the members' role in measure development.
Perinatal Care Measurement Advisory Panel

Eli Y. Adashi, MD, MS, CPE, FACOG, Brown University
Hani Atrash, MD, Health Resources and Services Administration
Sean Currigan, MPH, American College of Obstetricians and Gynecologist
Nicole Garro, MPH, March of Dimes
Rebekah E. Gee, MD, MPH, FACOG, Louisiana State University School of Public Health
Tina D. Groat, MD, MBA, FACOG, UnitedHealthcare
Renee Miskimmin, MD, MBA, FAAFP, Virginia Premier Health Plan
Catherine Ruhl, MS, CNM, Association of Women's Health, Obstetric and Neonatal Nurses
Carol Sakala, PhD, MSPH, National Partnership for Woman & Families
Carolyn L. Westhoff, MD, MSc, Planned Parenthood Federation of America

Measure Developer/Steward Updates and Ongoing Maintenance
Ad.2 **Year the measure was first released:** 1997
Ad.3 **Month and Year of most recent revision:** 07, 2014
Ad.4 **What is your frequency for review/update of this measure?** Approximately every 3 years, sooner if the clinical guidelines have changed significantly.
Ad.5 **When is the next scheduled review/update for this measure?** 12, 2017

Ad.7 **Disclaimers:** These performance measures are not clinical guidelines and do not establish a standard of medical care, and have not been tested for all potential applications.
THE MEASURES AND SPECIFICATIONS ARE PROVIDED "AS IS" WITHOUT WARRANTY OF ANY KIND.

Ad.8 **Additional Information/Comments:** NCQA Notice of Use. Broad public use and dissemination of these measures is encouraged and NCQA has agreed with NQF that noncommercial uses do not require the consent of the measure developer. Use by health care physicians in connection with their own practices is not commercial use. Commercial use of a measure requires the prior written consent of NCQA. As used herein, "commercial use" refers to any sale, license or distribution of a measure for commercial gain, or incorporation of a measure into any product or service that is sold, licensed or distributed for commercial gain, even if there is no actual charge for inclusion of the measure.
These performance measures were developed and are owned by NCQA. They are not clinical guidelines and do not establish a standard of medical care. NCQA makes no representations, warranties or endorsement about the quality of any organization or physician that uses or reports performance measures, and NCQA has no liability to anyone who relies on such measures. NCQA holds a copyright in these measures and can rescind or alter these measures at any time. Users of the measures shall not have the right to alter, enhance or otherwise modify the measures, and shall not disassemble, recompile or reverse engineer the source code or object code relating to the measures. Anyone desiring to use or reproduce the measures without modification for a noncommercial purpose may do so without obtaining approval from NCQA. All commercial uses must be approved by NCQA and are subject to a license at the discretion of NCQA.

MEASURE WORKSHEET

This document summarizes the evaluation of the measure as it progresses through NQF's Consensus Development Process (CDP). The information submitted by measure developers/stewards is included after the Brief Measure Information, Preliminary Analysis, and Pre-meeting Public and Member Comments sections.

To navigate the links in the worksheet: Ctrl + click link to go to the link; ALT + LEFT ARROW to return

Brief Measure Information

NQF #: 1731
Measure Title: PC-04 Health Care-Associated Bloodstream Infections in Newborns
Measure Steward: The Joint Commission
Brief Description of Measure: This measure assesses the number of staphylococcal and gram negative septicemias or bacteremias in high-risk newborns. This measure is a part of a set of five nationally implemented measures that address perinatal care (PC-01: Elective Delivery, PC-02: Cesarean Birth, PC-03: Antenatal Steroids, PC-05: Exclusive Breast Milk Feeding).
Developer Rationale: A health care-associated bloodstream infection in high-risk newborns remains a major patient safety concern. Effective preventive measures range from simple hand-washing protocols or closed medication delivery systems to more elaborate multidisciplinary quality improvement plans involving hand-washing, nutrition, skin care, respiratory care, vascular access, and diagnostic practices. Guidelines for the prevention of intravascular catheter-related infections are also available from the Centers for Disease Control and Prevention (CDC) to assist hospitals in establishing successful interventions to reduce the number of health care-associated bloodstream infections in newborns.
The measure will assist health care organizations (HCOs) to track evidence of a decrease in health care-associated bloodstream infections in newborns.

Numerator Statement: The outcome being measured is: Newborns with septicemia or bacteremia with ICD-10-CM Other Diagnosis Codes for newborn septicemia or bacteremia as defined in Appendix A, Table 11.10 with a Bloodstream Infection Confirmed OR ICD-10-CM Other Diagnosis Codes for sepsis as defined in Appendix A, Table 11.10.1 with a Bloodstream Infection Confirmed available at: http://manual.jointcommission.org/releases/TJC2015B2/

The only national hospital quality measure currently requiring patient-level risk adjustment is the Health Care-Associated Bloodstream Infections in Newborns (PC-04) outcome measure in the perinatal care measure set.
Denominator Statement: The outcome target population being measured is: Liveborn newborns with ICD-10-CM Other Diagnosis Codes for birth weight between 500 and 1499g as defined in Appendix A, Table 11.12, 11.13 or 11.14 OR Birth Weight between 500 and 1499g OR ICD-10-CM Other Diagnosis Codes for birth weight = > 1500g as defined in Appendix A, Table 11.15 or 11.16 OR Birth Weight = > 1500g who experienced one or more of the following:
o Experienced death
o ICD-10-PCS Principal Procedure Code or ICD-10-PCS Other Procedure Codes for major surgery as defined in Appendix A, Table 11.18
o ICD-10-PCS Principal Procedure Code or ICD-10-PCS Other Procedure Codes for mechanical ventilation as defined in Appendix A, Table 11.19
o Transferred in from another acute care hospital or health care setting within 2 days of birth.
Denominator Exclusions: • ICD-10-CM Principal Diagnosis Code for septicemias or bacteremias as defined in Appendix A, Table 11.10.2
• ICD-10-CM Other Diagnosis Codes for septicemias or bacteremias as defined in Appendix A, Table 11.10.2 or ICD-10-CM Principal or Other Diagnosis Codes for newborn septicemia or bacteremia as defined in Appendix A, Table 11.10 with a Bloodstream Infection Present on Admission
• ICD-10-CM Other Diagnosis Codes for birth weight < 500g as defined in Appendix A, Table 11.20 OR Birth Weight < 500g

• Length of Stay < 2 days
• Enrolled in clinical trials

Measure Type: Outcome
Data Source: Paper Medical Records
Level of Analysis: Facility, Population : National

IF Endorsement Maintenance – Original Endorsement Date: Apr 02, 2012 **Most Recent Endorsement Date:** Mar 30, 2012

Maintenance of Endorsement -- Preliminary Analysis

To maintain NQF endorsement endorsed measures are evaluated periodically to ensure that the measures still meets the NQF endorsement criteria ("maintenance"). The emphasis for maintaining endorsement is focused on how effective the measure is for promoting improvements in quality. Endorsed measures should have some experience from the field to inform the evaluation. The emphasis for maintaining endorsement is noted for each criterion.

Criteria 1: Importance to Measure and Report

1a. Evidence
Maintenance measures – less emphasis on evidence unless there is new information or change in evidence since the prior evaluation.

1a. Evidence. The evidence requirements for a health outcomes measure include providing rationale that supports the relationship of the health outcome to processes or structures of care. The guidance for evaluating the clinical evidence asks if the relationship between the measured health outcome and at least one clinical action is identified and supported by the stated rationale.

Summary of evidence:

- Health care-associated bacteremia is a significant problem for infants admitted into neonatal intensive care units (NICUs) and other hospital units. This is especially true for very low birth weight infants who are at high risk for these infections due to their immature immune systems and need for invasive monitoring and supportive care.
- Mortality rates are high and infections result in increased length of stay as well as increased hospital costs and charges.
- Studies demonstrate that educational interventions aimed at neonatal nurses show consistent findings of reductions in infection rates of 21-40%.

Changes to evidence from last review
☒ **The developer attests that there have been no changes in the evidence since the measure was last evaluated.**
☐ **The developer provided updated evidence for this measure.**

Exception to evidence NA

Questions for the Committee:
- *The developer attests the underlying evidence for the measure has not changed since the last NQF endorsement review. Does the Committee agree the evidence basis for the measure has not changed and there is no need for repeat discussion and voting on Evidence?*

Preliminary rating for evidence: ☒ **Pass** ☐ **No Pass**

1b. Gap in Care/Opportunity for Improvement and 1b. Disparities
Maintenance measures – increased emphasis on gap and variation

1b. Performance Gap. The performance gap requirements include demonstrating quality problems and opportunity for improvement.

- The developer reports "Health care-associated bloodstream infections continue to persist despite the fact that

standardized guidelines have been developed for intravascular catheter care. The goal of eliminating health care-associated bloodstream infections has not been met for all hospitals reporting the data."

- In January 2014, The Joint Commission required mandatory reporting of the PC measure set for all accredited hospitals with 1100 births or more annually and in January 2016 it was lowered again to all hospitals with more than 300 births annually.
- The 2014 mean rate of 2.9% remains above the target goal of 0%.

The developers provided the following data:

	2011	2012	2013	2014
# hospitals	109	89	122	1,218
# patients	6,490	2,570	4,861	71,676
National aggregate rate	0.09%	0.08%	2.5%	3.2%
10^{th}-25^{th} -75^{th} -90^{th} %tile	0 – 0 – 0 - 3.8%	0 – 0 – 0 - 2.2%	0 - 0 – 2% - 6.2%	0 – 0 – 1.8% - 7.1%
Mean hospital rate (SD)	0.04755 (0.19254)	0.01566 (0.10635)	0.02693 (0.10249)	0.02981 (0.09815)

Disparities

- The developer states that "There is no mention of disparities related to race or socioeconomic status regarding the incidence of health care associated bloodstream infections in newborns. Although the literature supports premature newborns with very low birth weight > 1500 g as the most vulnerable group of newborns susceptible to health care-associated bloodstream infections."

Questions for the Committee:

- *Is there a gap in care that warrants a national performance measure?*

Preliminary rating for opportunity for improvement: ☒ **High** ☐ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 1: Importance to Measure and Report (including 1a, 1b, 1c)

1a. Evidence to Support Measure Focus

Comments:

Meets criteria.

1b. Performance Gap

Comments:

Yes- continued variation in performance demonstrated. How can there be no disparity data on this with African American PTB being a known disparity- and this measure specific to this population?

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability

2a1. Reliability Specifications

Maintenance measures – no change in emphasis – specifications should be evaluated the same as with new measures

2a1. Specifications requires the measure, as specified, to produce consistent (reliable) and credible (valid) results about the quality of care when implemented.

Data source(s): Paper medical records; Vital Records reports, delivery logs and clinical information systems were added

Specifications:

- The specifications have been updated to ICD-10.
- A number of changes have been made to the numerator, denominator, and exclusions.

- The numerator is calculated with two data elements and is "Newborns with septicemia or bacteremia with ICD-10-CM Other Diagnosis Codes for newborn septicemia or bacteremia as defined in Appendix A, Table 11.10 with a Bloodstream Infection Confirmed OR ICD-10-CM Other Diagnosis Codes for sepsis as defined in Appendix A, Table 11.10.1 with a Bloodstream Infection Confirmed available at: http://manual.jointcommission.org/releases/TJC2015B2/"
- The denominator is calculated with 11 data elements and is "The outcome target population being measured is: Liveborn newborns with ICD-10-CM Other Diagnosis Codes for birth weight between 500 and 1499g as defined in Appendix A, Table 11.12, 11.13 or 11.14 OR Birth Weight between 500 and 1499g OR ICD-10-CM Other Diagnosis Codes for birth weight = > 1500g as defined in Appendix A, Table 11.15 or 11.16 OR Birth Weight = > 1500g who experienced one or more of the following:
  - Experienced death
  - ICD-10-PCS Principal Procedure Code or ICD-10-PCS Other Procedure Codes for major surgery as defined in Appendix A, Table 11.18
  - ICD-10-PCS Principal Procedure Code or ICD-10-PCS Other Procedure Codes for mechanical ventilation as defined in Appendix A, Table 11.19
  - Transferred in from another acute care hospital or health care setting within 2 days of birth."
- The measure is risk adjusted (more below).
- The calculation algorithm is included.
- Sampling is not allowed.

Questions for the Committee:

- *Are the changes to the numerator and denominator appropriate?*
- *Are all the data elements clearly defined? Are all appropriate codes included?*
- *Is the logic or calculation algorithm clear?*
- *Is it likely this measure can be consistently implemented?*

2a2. Reliability Testing Testing attachment
Maintenance measures – less emphasis if no new testing data provided

2a2. Reliability testing demonstrates if the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise enough to distinguish differences in performance across providers.

For maintenance measures, summarize the reliability testing from the prior review:

- The measure was tested using Inter-Rater Reliability (IRR) by the ORYX vendor, for 106 hospitals with 26,302 records. IRR is an appropriate method of assessing data element reliability for chart abstraction. The agreement rate for the data elements "admission date", "admission type", and "point of origin" was 99.85% for each; for data element "birth weight" agreement was 94.11%.

Describe any updates to testing
No updates to reliability testing

SUMMARY OF TESTING
Reliability testing level ☐ **Measure score** ☒ **Data element** ☐ **Both**
Reliability testing performed with the data source and level of analysis indicated for this measure ☒ **Yes** ☐ **No**

Method(s) of reliability testing Inter-rater reliability (Above)

Results of reliability testing The results of the reliability testing are within acceptable levels (above). Only %agreement was reported – no Kappa or ICC.

Guidance from the Reliability Algorithm

Precise specifications (Box 1) → Empirical reliability testing (Box 2) → Data element testing (Box 8) → appropriate method- IRR (Box 9) → high or moderate confidence of reliability of numerator data element → moderate (highest rating possible in the absence of testing the measure score)

Questions for the Committee:

- *The developer has not provided any additional testing for reliability since the last NQF endorsement review. Does the Committee agree there is no need for repeat discussion and voting on reliability?*

Preliminary rating for reliability: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

2b. Validity
Maintenance measures – less emphasis if no new testing data provided

2b1. Validity: Specifications

2b1. Validity Specifications. This section should determine if the measure specifications are consistent with the evidence.

Specifications consistent with evidence in 1a. ☒ **Yes** ☐ **Somewhat** ☐ **No**

Question for the Committee:

- *Are the specifications consistent with the evidence?*

2b2. Validity testing

2b2. Validity Testing should demonstrate the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality.

For maintenance measures, summarize the validity testing from the prior review:
Face validity only. Continued face validity has been determined through feedback from measure users.

Describe any updates to validity testing
Additional validity testing of the measure score was performed.

SUMMARY OF TESTING
Validity testing level ☒ **Measure score** ☐ **Data element testing against a gold standard** ☐ **Both**

Method of validity testing of the measure score:

- ☐ **Face validity only**
- ☒ **Empirical validity testing of the measure score**

Validity testing method:

- Data from Q3 and Q4, 2014, and Q1 and Q2, 2015, (1,345 hospitals submitted 2,695,467 inpatient records) for all the Joint Commission perinatal care measures. Measure convergent validity for PC-04 was assessed using hospitals patient level data from The Joint commission warehouse.
- Spearman rank order correlation was performed to correlate the performance across all perinatal care measures.

Validity testing results:

- The developer provides scatter plots and a correlation table. They note that the correlation on the performance between this measure and PC-02 (Cesarean birth) is relatively weak but statistically significant. The correlation with the other PC measures is not significant.

Questions for the Committee:

- *Do you expect a correlation between this measure and the other perinatal measures?*
- *Is the test sample adequate to generalize for widespread implementation?*
- *Do the results demonstrate sufficient validity so that conclusions about quality can be made?*
- *Do you agree that the score from this measure as specified is an indicator of quality?*

2b3-2b7. Threats to Validity

2b3. Exclusions:

There are five exclusions:

- ICD-10-CM Principal Diagnosis Code for septicemias or bacteremias as defined in Appendix A, Table 11.10.2
- ICD-10-CM Other Diagnosis Codes for septicemias or bacteremias as defined in Appendix A, Table 11.10.2 or ICD-10-CM Principal or Other Diagnosis Codes for newborn septicemia or bacteremia as defined in Appendix A, Table 11.10 with a Bloodstream Infection Present on Admission
- ICD-10-CM Other Diagnosis Codes for birth weight < 500g as defined in Appendix A, Table 11.20 OR Birth Weight < 500g
- Length of Stay < 2 days
- Enrolled in clinical trials

The developer notes "The difference between the 10th and 90th percentiles of the distribution of exclusion rates is narrow indicating that the occurrence is random and likely would not bias performance results." A list of rationales for each exclusion is included.

Questions for the Committee:

- *Are the exclusions consistent with the evidence?*
- *Are any patients or patient groups inappropriately excluded from the measure? Are there additional exclusions that should be included?*
- *Are the exclusions/exceptions of sufficient frequency and variation across providers to be needed (and outweigh the data collection burden)?*

2b4. Risk adjustment: **Risk-adjustment method** ☐ **None** ☒ **Statistical model** ☐ **Stratification**

- The developers note that "the AHRQ Pediatric Quality indicators, NQI #3 – Neonatal Bloodstream Infection, was a similar measure already developed with a tested risk model. We initially used all the risk factors identified for the AHRQ risk model. "
- The measure adjusts for six risk factors, which were all deemed significant at p<0.0001:
  - Birth Weight 500 to 749g
  - Birth Weight 750 to 750g
  - Modified DRG Newborn Transfers Out or Died
  - Congenital Anomaly Gastrointestinal Anomaly
  - Congenital Anomaly Cardiovascular Anomaly
  - Out-born Birth Newborns Transfers In

Conceptual rationale for SDS factors included? ☒ **Yes** ☐ **No**

- Race and ethnicity data were the SDS factors available in the data for testing.

SDS factors included in risk model? ☐ **Yes** ☒ **No**

- African American race and Hispanic ethnicity were both found to be statistically significant in the model, after

adjusting for the other risk factors, and increased the c-statistic of the risk model to 0.702.

Risk adjustment summary

- Risk model discrimination statistics and calibration curves were used to validate the adequacy of the risk model.
- The c-statistic for the final model was c=0.654**.**

Questions for the Committee:

- *Is an appropriate risk-adjustment strategy included in the measure?*
- *Are the candidate and final variables included in the risk adjustment model adequately described for the measure to be implemented?*
- *Are all of the risk adjustment variables present at the start of care? If not, describe the rationale provided.*

2b5. Meaningful difference *(can statistically significant and clinically/practically meaningful differences in performance measure scores can be identified)*:

The developer reports the following descriptive statistics for PC-04 measure:
N=523 hospitals n = 469,910

Min = 0%
Mean: 4.9%
Percentile 10%: 0%
Percentile 25%: 0%
Median: 2.2%
Percentile 75%: 7.7%
Percentile 90%: 13.9%
Max = 66.1%

The Joint Commission's Target Analysis uses two methods: Given the national average for a performance measure, a target range is constructed. Using generalized linear mixed models methodology (also known as hierarchical models), a predicted estimate of an HCO's performance, with a corresponding 95% confidence interval, is generated. This confidence interval is compared to the target range, to determine the HCOs' rating. The estimate of the organization's true performance is based on both the data from that organization and on data from the entire set of reporting organizations. The object of target analysis is to compare a health care organizations (HCO) data against a comparative norm for the purpose of evaluating performance improvement opportunities. When an organization's performance level is statistically significantly different from a comparative norm, it is considered a statistical deviation. A statistical deviation may be desirable or undesirable depending on the "direction of improvement" of the measure.

PC-02 Distribution of Outliers
2011 1st Quarter Data:
Scores on this measure: N=79, Mean 1.78%, SD 0.1134
10th Percentile= 0%
25th Percentile= 0%
50th Percentile= 0%
75th Percentile= 0%
90th Percentile= 1.64%
79 (100%) Neutral – results not significantly different from target range

Question for the Committee:

- *Does this measure identify meaningful differences about quality?*

2b6. Comparability of data sources/methods:

N/A

2b7. Missing Data

- The developer states "Any file with missing data will result in a measure category assignment of X and rejection of the file from the warehouse, unless the data are not used to process the measure. If the data are used to process the measure and have been reported by the abstractor as No/UTD, the case will result in a measure category assignment of E (i.e., failed measure, not rejected)."

Guidance from algorithm: Specifications consistent with evidence (Box 1) → Assessed all potential threats to validity (Box 2) **Meaningful differences may be an issue here** → empirical testing (Box 3) →testing of measure score (Box 6) →appropriate testing method (Box 7) → testing results moderate → moderate

\

Preliminary rating for validity: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 2: Scientific Acceptability of Measure Properties (including all 2a, 2b, and 2d)

2a1. & 2b1. Specifications

Comments:

Changes to numerator and denominator seem appropriate. Data elements are clearly defined algorithm is clear. Risk adjustments seem appropriate consistent implementation is likely.

Empirical validity is demonstrated. Correlation to other PC measures: weak correlation with cesarean, other PC correlation not significant- perhaps it would correlate with other quality measures outside of PC.

2a2. Reliability Testing

Comments:

Supports moderate preliminary rating

2b2. Validity Testing

Comments:

Supports moderate validity rating

2b3. Exclusions Analysis

2b4. Risk Adjustment/Stratification for Outcome or Resource Use Measures

2b5. Identification of Statistically Significant & Meaningful Differences In Performance

2b6. Comparability of Performance Scores When More Than One Set of Specifications

2b7. Missing Data Analysis and Minimizing Bias

Comments:

Appropriate exclusions. Appropriate risk adjustment. Demonstrates meaningful differences. Sppropriate treatment of missing data.

Criterion 3. Feasibility

Maintenance measures – no change in emphasis – implementation issues may be more prominent

3. Feasibility is the extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

- The developer states "At the present time, hospitals using this performance measure generally collect measure data via manual review of the paper medical record, the EMR or a combination of both. Collected data are submitted to The Joint Commission on a quarterly basis, by way of contracted performance measurement system vendors"
- The specifications are freely available and there are no fees or licensing requirements.
- The developer is currently retooling the measure for capture from electronic sources and plans to test in 2016.

Questions for the Committee:

- *Are the required data elements routinely generated and used during care delivery?*
- *Are the required data elements available in electronic form, e.g., EHR or other electronic sources?*

○ *Is the data collection strategy ready to be put into operational use?*

Preliminary rating for feasibility: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 3: Feasibility

3a. Byproduct of Care Processes
3b. Electronic Sources
3c. Data Collection Strategy

Comments:

The required data elements are generated during care in electronic data sources and are ready for operational use.

Criterion 4: Usability and Use
Maintenance measures – increased emphasis – much greater focus on measure use and usefulness, including both impact /improvement and unintended consequences

4. Usability and Use evaluate the extent to which audiences (e.g., consumers, purchasers, providers, policymakers) use or could use performance results for both accountability and performance improvement activities.

Current uses of the measure
Publicly reported? ☒ **Yes** ☐ **No**

Current use in an accountability program? ☒ **Yes** ☐ **No**
OR
Planned use in an accountability program? ☐ **Yes** ☐ **No**

Accountability program details

- Public Reporting --Quality Check® http://www.qualitycheck.org/consumer/searchQCR.aspx
- Regulatory and Accreditation Programs Hospital Accreditation Program http://jointcommission.org

Improvement results

- The developer reports some progress has been made in eliminating health care-associated bloodstream infections in newborns. In 2014, hospitals in in the median, lower quartile and 10th percentile reported no infections.
- There was a spike in the aggregate rate (to 3.2% from 0.02%) when 1060 more hospitals were added in 2014. In January 2016 an additional 821 hospitals will begin reporting, which may cause another increase in rates, but with these new facilities 80% of the accredited hospitals with maternity services will be included.

Unexpected findings (positive or negative) during implementation

The developer notes four unintended consequences; steps were taken to address each. These included:

- Increase in burden of data abstraction – ACTION: Vital Records reports, delivery logs and clinical information systems were added as acceptable data sources
- Cases with a length of stay greater than 120 days were being inappropriately excluded – ACTION: This exclusion was removed.
- Cases with infection codes were failing when bloodstream infections were present on admission. ACTION: New

data element added.

- Cases coded with infections were failing when newborns experienced bloodstream infections that were not health care-associated later during the hospitalization. ACTION- new data element added.

Potential harms None reported

Feedback:

***Questions for the Committee**:*

- *Committee members should share any experience with use of this measure.*
- *What does experience with this measure tell various stakeholders?*
- *How can the performance results be used to further the goal of high-quality, efficient healthcare?*
- *Do the benefits of the measure outweigh any potential unintended consequences?*

Preliminary rating for usability and use: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 4: Usability and Use

4a. Accountability and Transparency
4b. Improvement
4c. Unintended Consequences
Comments:
Support moderate rating.

Criterion 5: Related and Competing Measures

Related or competing measures

- 0304: Late sepsis or meningitis in Very Low Birth Weight (VLBW) neonates (risk-adjusted)
- 0478: Neonatal Blood Stream Infection Rate (NQI 03)

Harmonization

- The developer states that this measure has been harmonized to the extent possible with 0478; "however, there are intrinsic differences which are addressed in a comparison table in the attachment found in Section A.1 Supplemental Materials."

Pre-meeting public and member comments

-

NQF #: 1731 **NQF Project:** Perinatal and Reproductive Health Project

1. IMPACT, OPPORTUITY, EVIDENCE - IMPORTANCE TO MEASURE AND REPORT

Importance to Measure and Report is a threshold criterion that must be met in order to recommend a measure for endorsement. All three subcriteria must be met to pass this criterion. See guidance on evidence.
Measures must be judged to be important to measure and report in order to be evaluated against the remaining criteria.
(evaluation criteria)

1c.1 **Structure-Process-Outcome Relationship** *(Briefly state the measure focus, e.g., health outcome, intermediate clinical outcome, process, structure; then identify the appropriate links, e.g., structure-process-health outcome; process- health outcome; intermediate clinical outcome-health outcome)*:
The focus of the measure is to prevent health care associated bloodstream infections in newborns. Hospital develops an infection prevention program for newborns >> population determined >> population assessed >> prevention measures instituted >> no evidence of a health care associated bloodstream infection while in the hospital >> patient discharged >> no bloodstream infection detected related to hospitalization >> reduced morbidity/mortality related to blood stream infections.

1c.2-3 **Type of Evidence** *(Check all that apply)*:
Clinical Practice Guideline
Systematic review of body of evidence (other than within guideline development)

1c.4 **Directness of Evidence to the Specified Measure** *(State the central topic, population, and outcomes addressed in the body of evidence and identify any differences from the measure focus and measure target population)*:
The central topic for the measure is the prevention of health care-associated bloodstream infections in newborns. The majority of high-risk newborns are at risk for health care-associated bloodstream infections. The target population for the performance measure is consistent with the body of evidence supporting the need for a health care-associated bloodstream infection prevention program.

1c.5 **Quantity of Studies in the Body of Evidence** *(Total number of studies, not articles)*: There have been numerous cohort studies and surveillance surveys conducted since 1996 through 2009 documenting the prevalence of health care-associated bloodstream infections in newborns and the associated risk factors. Additionally, four databases were searched: PubMed, CINAHL, Cochrane and OVID to examine how educational interventions could help neonatal nurses reduce infection rates in patients with central venous catheters. Ten studies were identified which measured rates before and after interventions.

The Cochrane Collaboration also reviewed eight randomized control trials regarding the use of prophylactic Vancomycin and other systemic antibiotics as other interventions to reduce the incidence of health care-associated bloodstream infections. Several observational studies were reviewed comparing the early removal of central venous catheters versus expectant management.

1c.6 **Quality of Body of Evidence** (*Summarize the certainty or confidence in the estimates of benefits and harms to patients across studies in the body of evidence resulting from study factors. Please address: a) study design/flaws; b) directness/indirectness of the evidence to this measure (e.g., interventions, comparisons, outcomes assessed, population included in the evidence); and c) imprecision/wide confidence intervals due to few patients or events):* The quality of evidence supporting educational interventions aimed at neonatal nurses is quite high with studies published that have involved patients in neonatal intensive care units (NICUs). As noted in the CDC guidelines, standardization of aseptic care can reduce adverse patient outcomes while at the same time decreasing overall costs. Nine studies examining educational interventions aimed at neonatal nurses resulted in a 40% reduction in catheter-related bloodstream infections with eight demonstrating statistically significant reductions. No study design flaws were identified during the review.

1c.7 **Consistency of Results across Studies** *(Summarize the consistency of the magnitude and direction of the effect):* Results of studies evaluating educational interventions aimed at neonatal nurses show consistent findings of reductions in event rates of 21-40%

in 10 studies. Several quasi-experimental studies have demonstrated that NICUs can lower their infection rates (based on positive blood cultures) from as high as 13.5 per 1,000 patient days to as low as 3.0 per 1,000 patient days. Reports spanning the past four decades have consistently demonstrated that risk for infection declines following standardization of aseptic care.

1c.8 **Net Benefit** *(Provide estimates of effect for benefit/outcome; identify harms addressed and estimates of effect; and net benefit - benefit over harms)***:**
As described before, educational interventions aimed at neonatal nurses show consistent findings of reductions in event rates of 21-40%.There is no evidence that standardized aseptic care, which is often a key component of nursing educational interventions, results in harm to patients.

1c.9 **Grading of Strength/Quality of the Body of Evidence**. Has the body of evidence been graded? No

1c.10 **If body of evidence graded, identify the entity that graded the evidence including balance of representation and any disclosures regarding bias:**

1c.11 **System Used for Grading the Body of Evidence:** Other

1c.12 **If other, identify and describe the grading scale with definitions:** Although grading of the evidence was not determined during our systematic review, it was determined that the guideline developers accounted for a balanced representation of information, looked beyond one specialty group or discipline, and provided information that was accessible and met the requirements set out in this measure maintenance form.

1c.13 **Grade Assigned to the Body of Evidence:** Not Applicable

1c.14 **Summary of Controversy/Contradictory Evidence:** The use of Vancomycin and other systemic antibiotics and early removal of central venous catheters to reduce blood stream infections are controversial. Low dose Vancomycin and other systemic antibiotics reduce the incidence of health care-associated bloodstream infections, but there is concern regarding routine use and the development of resistant organisms. The early removal of central venous catheters may also be of benefit; however, there are no randomized control trials to validate this practice.

1c.15 **Citations for Evidence other than Guidelines*(Guidelines addressed below)*:**

- Adams-Chapman, I. & Stoll, B.J. (2002). Prevention of nosocomial infections in the neonatal intensive care unit. Current Opinion in Pediatrics.14 (2):157-64.
- Aly, H., Herson, V., Duncan, A., et al. (2005). Is bloodstream infection preventable among premature infants? A tale of two cities. Pediatrics. 115(6):1513-8.
- Bloom, B.T., Craddock, A., Delmore, P.M., et al. (2003). Reducing acquired infections in the NICU: observing and implementing meaningful differences in process between high and low acquired infection rate centers. Journal of Perinatology. 23(6):489-92.
- Clark, R., Powers, R., White, R., Bloom, B., Sanchez, P., & Benjamin, D.K., Jr. (2004a). Prevention and treatment of nosocomial sepsis in the NICU. Journal of Perinatology. 4; 24(7):446-53.
- Clark, R., Powers, R., White, R., Bloom, B., Sanchez, P., & Benjamin, D.K., Jr. (2004b). Nosocomial infection in the NICU: a medical complication or unavoidable problem? Journal of Perinatology. 24(6):382-8.
- Craft, A.P., Finer, N. & Barrington, K.J. (2009). Vancomycin for prophylaxis against sepsis in preterm neonates (Review). The Cochrane Collaboration. Issue 1.
- Horbar, J.D., Rogowski, J., Plsek, P.E., et al. (2001). Collaborative quality improvement for neonatal intensive care. NIC/Q Project Investigators of the Vermont Oxford Network. Pediatrics. 107(1):14-22.
- Jardine, L.A., Inglis, G.D.T. & Davies, M.W. (2010). Prophylactic systemic antibiotics to erduce morbidity and mortality in neonates with central venous catheters (Review). The Cochrane Collaboration. Issue 5.
- Kilbride, H.W., Wirtschafter, D.D., Powers, R.J., & Sheehan, M.B. (2003a). Implementation of evidence-based potentially better practices to decrease nosocomial infections. Pediatrics. 111(4 Pt 2):e519-33.
- Kilbride, H.W., Powers, R., Wirtschafter, D.D., et al. (2003b). Evaluation and development of potentially better practices to prevent neonatal nosocomial bacteremia. Pediatrics. 111(4 Pt 2):e504-18.
- Lam, B.C., Lee, J., & Lau, Y.L. (2004). Hand Hygiene Practices in a Neonatal Intensive Care Unit: A Multimodal Intervention and Impact on Nosocomial Infection. Pediatrics.114 (5):e565.
- Ng, P.C., Wong, H.L., Lyon, D.J., et al. (2004). Combined use of alcohol hand rub and gloves reduces the incidence of late

onset infection in very low birthweight infants. Archives of Disease in Childhood Fetal & Neonatal Edition. 89(4):F336-40.
• Schelonka, R.L., Scruggs, S., Nichols, K., Dimmitt, R.A., & Carlo, W.A. (2006). Sustained reductions in neonatal nosocomial infection rates following a comprehensive infection control intervention. Journal of Perinatology. 26(3):176-9.
• Semelsberger, C.F. (2009). Educational interventions to reduce the rate of catheter-related bloodstream infections in the NICU: a review of the research literature. Neonatal Netw. 28(6) 391-5.
• Vasudevan, C. & McGuire, W. (2011). Early removal versus expectant management of central venous catheters in neonates with bloodstream infection. (2111). The Cochrane Collaboration. Issue 8.

1c.16 Quote verbatim, <u>the specific guideline recommendation</u> *(Including guideline # and/or page #)*:

Summary of Recommendations

Education, Training and Staffing

1. Educate healthcare personnel regarding the indications for intravascular catheter use, proper procedures for the insertion and maintenance of intravascular catheters, and appropriate infection control measures to prevent intravascular catheter-related infections [7–15]. Category IA
2. Periodically assess knowledge of and adherence to guidelines for all personnel involved in the insertion and maintenance of intravascular catheters [7–15]. Category IA
3. Designate only trained personnel who demonstrate competence for the insertion and maintenance of peripheral and central intravascular catheters. [14–28]. Category IA

4. Ensure appropriate nursing staff levels in ICUs. Observational studies suggest that a higher proportion of "pool nurses" or an elevated patient–to-nurse ratio is associated with CRBSI in ICUs where nurses are managing patients with CVCs [29–31]. Category IB

Selection of Catheters and Sites Peripheral Catheters and Midline Catheters

1. In adults, use an upper-extremity site for catheter insertion. Replace a catheter inserted in a lower extremity site to an upper extremity site as soon as possible. Category II
2. In pediatric patients, the upper or lower extremities or the scalp (in neonates or young infants) can be used as the catheter insertion site [32, 33]. Category II
3. Select catheters on the basis of the intended purpose and duration of use, known infectious and non-infectious complications (e.g., phlebitis and infiltration), and experience of individual catheter operators [33–35]. Category IB
4. Avoid the use of steel needles for the administration of fluids and medication that might cause tissue necrosis if extravasation occurs [33, 34]. Category IA
5. Use a midline catheter or peripherally inserted central catheter (PICC), instead of a short peripheral catheter, when the duration of IV therapy will likely exceed six days. Category II
6. Evaluate the catheter insertion site daily by palpation through the dressing to discern tenderness and by inspection if a transparent dressing is in use. Gauze and opaque dressings should not be removed if the patient has no clinical signs of infection. If the patient has local tenderness or other signs of possible CRBSI, an opaque dressing should be removed and the site inspected visually. Category II
7. Remove peripheral venous catheters if the patients develop signs of phlebitis (warmth, tenderness, erythema or palpable venous cord), infection, or a malfunctioning catheter [36]. Category IB

Central Venous Catheters

1. Weigh the risks and benefits of placing a central venous device at a recommended site to reduce infectious complications against the risk for mechanical complications (e.g., pneumothorax, subclavian artery puncture, subclavian vein laceration, subclavian vein stenosis, hemothorax, thrombosis, air embolism, and catheter misplacement) [37–53]. Category IA
2. Avoid using the femoral vein for central venous access in adult patients [38, 50, 51, 54]. Category 1A
3. Use a subclavian site, rather than a jugular or a femoral site, in adult patients to minimize infection risk for nontunneled CVC placement [50–52]. Category IB
4. No recommendation can be made for a preferred site of insertion to minimize infection risk for a tunneled CVC. Unresolved issue
5. Avoid the subclavian site in hemodialysis patients and patients with advanced kidney disease, to avoid subclavian vein stenosis [53,55–58]. Category IA
6. Use a fistula or graft in patients with chronic renal failure instead of a CVC for permanent access for dialysis [59]. Category 1A
7. Use ultrasound guidance to place central venous catheters (if this technology is available) to reduce the number of cannulation attempts and mechanical complications. Ultrasound guidance should only be used by those fully trained in its technique. [60–64]. Category 1B
8. Use a CVC with the minimum number of ports or lumens essential for the management of the patient [65–68]. Category IB
9. No recommendation can be made regarding the use of a designated lumen for parenteral nutrition. Unresolved issue
10. Promptly remove any intravascular catheter that is no longer essential [69–72]. Category IA

11. When adherence to aseptic technique cannot be ensured (i.e catheters inserted during a medical emergency), replace the catheter as soon as possible, i.e, within 48 hours [37,73–76]. Category IB

Hand Hygiene and Aseptic Technique

1. Perform hand hygiene procedures, either by washing hands with conventional soap and water or with alcohol-based hand rubs (ABHR). Hand hygiene should be performed before and after palpating catheter insertion sites as well as before and after inserting, replacing, accessing, repairing, or dressing an intravascular catheter. Palpation of the insertion site should not be performed after the application of antiseptic, unless aseptic technique is maintained [12, 77–79]. Category IB
2. Maintain aseptic technique for the insertion and care of intravascular catheters [37, 73, 74, 76]. Category IB
3. Wear clean gloves, rather than sterile gloves, for the insertion of peripheral intravascular catheters, if the access site is not touched after the application of skin antiseptics. Category IC
4. Sterile gloves should be worn for the insertion of arterial, central, and midline catheters [37, 73, 74, 76]. Category IA
5. Use new sterile gloves before handling the new catheter when guidewire exchanges are performed. Category II
6. Wear either clean or sterile gloves when changing the dressing on intravascular catheters. Category IC

Maximal Sterile Barrier Precautions

1. Use maximal sterile barrier precautions, including the use of a cap, mask, sterile gown, sterile gloves, and a sterile full body drape, for the insertion of CVCs, PICCs, or guidewire exchange [14, 75, 76, 80]. Category IB
2. Use a sterile sleeve to protect pulmonary artery catheters during insertion [81]. Category IB

Skin Preparation

1. Prepare clean skin with an antiseptic (70% alcohol, tincture of iodine, or alcoholic chlorhexidine gluconate solution) before peripheral venous catheter insertion [82]. Category IB
2. Prepare clean skin with a >0.5% chlorhexidine preparation with alcohol before central venous catheter and peripheral arterial catheter insertion and during dressing changes. If there is a contraindication to chlorhexidine, tincture of iodine, an iodophor, or 70% alcohol can be used as alternatives [82, 83]. Category IA
3. No comparison has been made between using chlorhexidine preparations with alcohol and povidone-iodine in alcohol to prepare clean skin. Unresolved issue.
4. No recommendation can be made for the safety or efficacy of chlorhexidine in infants aged <2 months. Unresolved issue
5. Antiseptics should be allowed to dry according to the manufacturer's recommendation prior to placing the catheter [82, 83]. Category IB

Catheter Site Dressing Regimens

1. Use either sterile gauze or sterile, transparent, semipermeable dressing to cover the catheter site [84–87]. Category IA
2. If the patient is diaphoretic or if the site is bleeding or oozing, use a gauze dressing until this is resolved [84–87]. Category II
3. Replace catheter site dressing if the dressing becomes damp, loosened, or visibly soiled [84, 85]. Category IB
4. Do not use topical antibiotic ointment or creams on insertion sites, except for dialysis catheters, because of their potential to promote fungal infections and antimicrobial resistance [88, 89]. Category IB
5. Do not submerge the catheter or catheter site in water. Showering should be permitted if precautions can be taken to reduce the likelihood of introducing organisms into the
Guidelines for the Prevention of Intravascular Catheter-Related Infections catheter (e.g., if the catheter and connecting device are protected with an impermeable cover during the shower) [90–92]. Category IB
6. Replace dressings used on short-term CVC sites every 2 days for gauze dressings. Category II
7. Replace dressings used on short-term CVC sites at least every 7 days for transparent dressings, except in those pediatric patients in which the risk for dislodging the catheter may outweigh the benefit of changing the dressing [87, 93]. Category IB
8. Replace transparent dressings used on tunneled or implanted CVC sites no more than once per week (unless the dressing is soiled or loose), until the insertion site has healed. Category II
9. No recommendation can be made regarding the necessity for any dressing on well-healed exit sites of long-term cuffed and tunneled CVCs. Unresolved issue
10. Ensure that catheter site care is compatible with the catheter material [94, 95]. Category IB
11. Use a sterile sleeve for all pulmonary artery catheters [81]. Category IB
12. Use a chlorhexidine-impregnated sponge dressing for temporary short-term catheters in patients older than 2 months of age if the CLABSI rate is not decreasing despite adherence to basic prevention measures, including education and training, appropriate use of chlorhexidine for skin antisepsis, and MSB [93, 96–98]. Category 1B
13. No recommendation is made for other types of chlorhexidine dressings. Unresolved issue

14. Monitor the catheter sites visually when changing the dressing or by palpation through an intact dressing on a regular basis, depending on the clinical situation of the individual patient. If patients have tenderness at the insertion site, fever without obvious source, or other manifestations suggesting local or bloodstream infection, the dressing should be removed to allow thorough examination of the site [99–101]. Category IB
15. Encourage patients to report any changes in their catheter site or any new discomfort to their provider. Category II

Patient Cleansing
Use a 2% chlorhexidine wash for daily skin cleansing to reduce CRBSI [102–104]. Category II
Catheter Securement Devices Use a sutureless securement device to reduce the risk of infection for intravascular catheters [105]. Category II

Antimicrobial/Antiseptic Impregnated Catheters and Cuffs
Use a chlorhexidine/silver sulfadiazine or minocycline/rifampin -impregnated CVC in patients whose catheter is expected to remain in place >5 days if, after successful implementation of a comprehensive strategy to reduce rates of CLABSI, the CLABSI rate is not decreasing. The comprehensive strategy should include at least the following three components: educating persons, who insert and maintain catheters, use of maximal sterile barrier precautions, and a >0.5% chlorhexidine preparation with alcohol for skin antisepsis during CVC insertion [106–113]. Category IA

Systemic Antibiotic Prophylaxis
Do not administer systemic antimicrobial prophylaxis routinely before insertion or during use of an intravascular catheter to prevent catheter colonization or CRBSI [114]. Category IB

Antibiotic/Antiseptic Ointments
Use povidone iodine antiseptic ointment or bacitracin/gramicidin/ polymyxin B ointment at the hemodialysis catheter exit site after catheter insertion and at the end of each dialysis session only if this ointment does not interact with the material of the hemodialysis catheter per manufacturer's recommendation *59, 115–119]. Category IB

Antibiotic Lock Prophylaxis, Antimicrobial Catheter Flush and Catheter Lock Prophylaxis
Use prophylactic antimicrobial lock solution in patients with long term catheters who have a history of multiple CRBSI despite optimal maximal adherence to aseptic technique [120– 138]. Category II

Anticoagulants
Do not routinely use anticoagulant therapy to reduce the risk of catheter-related infection in general patient populations [139]. Category II

Replacement of Peripheral and Midline Catheters
1. There is no need to replace peripheral catheters more frequently than every 72-96 hours to reduce risk of infection and phlebitis in adults [36, 140, 141]. Category 1B
2. No recommendation is made regarding replacement of peripheral catheters in adults only when clinically indicated [142–144]. Unresolved issue
3. Replace peripheral catheters in children only when clinically indicated [32, 33]. Category 1B
4. Replace midline catheters only when there is a specific indication. Category II
Replacement of CVCs, Including PICCs and Hemodialysis Catheters
1. Do not routinely replace CVCs, PICCs, hemodialysis catheters, or pulmonary artery catheters to prevent catheter-related infections. Category IB
2. Do not remove CVCs or PICCs on the basis of fever alone. Use clinical judgment regarding the appropriateness of removing the catheter if infection is evidenced elsewhere or if a noninfectious cause of fever is suspected. Category II
3. Do not use guidewire exchanges routinely for non-tunneled catheters to prevent infection. Category IB
4. Do not use guidewire exchanges to replace a non-tunneled catheter suspected of infection. Category IB
5. Use a guidewire exchange to replace a malfunctioning non-tunneled catheter if no evidence of infection is present. Category IB
6. Use new sterile gloves before handling the new catheter when guidewire exchanges are performed. Category II

Umbilical Catheters
1. Remove and do not replace umbilical artery catheters if any signs of CRBSI, vascular insufficiency in the lower extremities, or thrombosis are present [145]. Category II

2. Remove and do not replace umbilical venous catheters if any signs of CRBSI or thrombosis are present [145]. Category II
3. No recommendation can be made regarding attempts to salvage an umbilical catheter by administering antibiotic treatment through the catheter. Unresolved issue
4. Cleanse the umbilical insertion site with an antiseptic before catheter insertion. Avoid tincture of iodine because of the potential effect on the neonatal thyroid. Other iodine-containing products (e.g., povidone iodine) can be used [146– 150]. Category IB
5. Do not use topical antibiotic ointment or creams on umbilical catheter insertion sites because of the potential to promote fungal infections and antimicrobial resistance [88, 89]. Category IA
6. Add low-doses of heparin (0.25—1.0 U/ml) to the fluid infused through umbilical arterial catheters [151–153]. Category IB
7. Remove umbilical catheters as soon as possible when no longer needed or when any sign of vascular insufficiency to the lower extremities is observed. Optimally, umbilical artery catheters should not be left in place >5 days [145, 154]. Category II
8. Umbilical venous catheters should be removed as soon as possible when no longer needed, but can be used up to 14 days if managed aseptically [155, 156]. Category II
9. An umbilical catheter may be replaced if it is malfunctioning, and there is no other indication for catheter removal, and the total duration of catheterization has not exceeded 5 days for an umbilical artery catheter or 14 days for an umbilical vein catheter. Category II

Peripheral Arterial Catheters and Pressure Monitoring Devices for Adult and Pediatric Patients
1. In adults, use of the radial, brachial or dorsalis pedis sites is preferred over the femoral or axillary sites of insertion to reduce the risk of infection [46, 47, 157, 158]. Category IB
2. In children, the brachial site should not be used. The radial, dorsalis pedis, and posterior tibial sites are preferred over the femoral or axillary sites of insertion [46]. Category II
3. A minimum of a cap, mask, sterile gloves and a small sterile fenestrated drape should be used during peripheral arterial catheter insertion [47, 158, 159]. Category IB
4. During axillary or femoral artery catheter insertion, maximal sterile barriers precautions should be used. Category II
5. Replace arterial catheters only when there is a clinical indication. Category II
6. Remove the arterial catheter as soon as it is no longer needed. Category II
7. Use disposable, rather than reusable, transducer assemblies when possible [160–164]. Category IB
8. Do not routinely replace arterial catheters to prevent catheter-related infections [165, 166, 167, 168]. Category II
9. Replace disposable or reusable transducers at 96-hour intervals. Replace other components of the system (including the tubing, continuous-flush device, and flush solution) at the time the transducer is replaced [37, 161]. Category IB
10. Keep all components of the pressure monitoring system (including calibration devices and flush solution) sterile [160, 169–171]. Category IA
11. Minimize the number of manipulations of and entries into the pressure monitoring system. Use a closed flush system (i.e, continuous flush), rather than an open system (i.e, one that requires a syringe and stopcock), to maintain the patency of the pressure monitoring catheters [163, 172]. Category II
12. When the pressure monitoring system is accessed through a diaphragm, rather than a stopcock, scrub the diaphragm with an appropriate antiseptic before accessing the system [163]. Category IA
Guidelines for the Prevention of Intravascular Catheter-Related Infections
19
13. Do not administer dextrose-containing solutions or parenteral nutrition fluids through the pressure monitoring circuit [163, 173, 174]. Category IA
14. Sterilize reusable transducers according to the manufacturers' instructions if the use of disposable transducers is not feasible [163, 173–176]. Category IA

Replacement of Administration Sets
1. In patients not receiving blood, blood products or fat emulsions, replace administration sets that are continuously used, including secondary sets and add-on devices, no more frequently than at 96-hour intervals, [177] but at least every 7 days [178–181]. Category IA
2. No recommendation can be made regarding the frequency for replacing intermittently used administration sets. Unresolved issue
3. No recommendation can be made regarding the frequency for replacing needles to access implantable ports. Unresolved issue
4. Replace tubing used to administer blood, blood products, or fat emulsions (those combined with amino acids and glucose in a 3-in-1 admixture or infused separately) within 24 hours of initiating the infusion [182–185]. Category IB
5. Replace tubing used to administer propofol infusions every 6 or 12 hours, when the vial is changed, per the manufacturer's recommendation (FDA website Medwatch) *186+. Category IA
6. No recommendation can be made regarding the length of time a needle used to access implanted ports can remain in place.

Unresolved issue

Needleless Intravascular Catheter Systems
1. Change the needleless components at least as frequently as the administration set. There is no benefit to changing these more frequently than every 72 hours. [39, 187–193]. Category II
Guidelines for the Prevention of Intravascular Catheter-Related Infections
20
2. Change needleless connectors no more frequently than every 72 hours or according to manufacturers' recommendations for the purpose of reducing infection rates *187, 189, 192, 193]. Category II
3. Ensure that all components of the system are compatible to minimize leaks and breaks in the system [194]. Category II
4. Minimize contamination risk by scrubbing the access port with an appropriate antiseptic (chlorhexidine, povidone iodine, an iodophor, or 70% alcohol) and accessing the port only with sterile devices [189, 192, 194–196]. Category IA
5. Use a needleless system to access IV tubing. Category IC
6. When needleless systems are used, a split septum valve may be preferred over some mechanical valves due to increased risk of infection with the mechanical valves [197–200]. Category II

Performance Improvement
Use hospital-specific or collaborative-based performance improvement initiatives in which multifaceted strategies are "bundled" together to improve compliance with evidence-based recommended practices [15, 69, 70, 201–205]. Category IB

1c.17 **Clinical Practice Guideline Citation:** • O´Grady, N.P., Alexander, M., Burns, L.A., Dellinger, P., Garland, J., Heard, S.O., Lipsett, P.A. et al. (2011). Guidelines for the prevention of intravascular catheter-related infections, 2011. Centers for Disease Control and Prevention. Retrieved August 9, 2011 at: http://www.cdc.gov/hicpac/pdf/guidelines/bsi-guidelines-2011.pdf

1c.18 **National Guideline Clearinghouse or other URL:** http://www.cdc.gov/hicpac/pdf/guidelines/bsi-guidelines-2011.pdf

1c.19 **Grading of Strength of Guideline Recommendation.** Has the recommendation been graded? Yes

1c.20 **If guideline recommendation graded, identify the entity that graded the evidence including balance of representation and any disclosures regarding bias:** The Centers for Disease Control and Prevention

1c.21 **System Used for Grading the Strength of Guideline Recommendation:** Other

1c.22 **If other, identify and describe the grading scale with definitions:** The system for categorizing recommendations in this guideline is as follows:

Category IA. Strongly recommended for implementation and strongly supported by well-designed experimental, clinical, or epidemiologic studies.

Category IB. Strongly recommended for implementation and supported by some experimental, clinical, or epidemiologic studies and a strong theoretical rationale; or an accepted practice (e.g., aseptic technique) supported by limited evidence.

Category IC. Required by state or federal regulations, rules, or standards.

Category II. Suggested for implementation and supported by suggestive clinical or epidemiologic studies or a theoretical rationale. Unresolved issue. Represents an unresolved issue for which evidence is insufficient or no consensus regarding efficacy exists.

1c.23 **Grade Assigned to the Recommendation:** Varies with the majority of recommendations Level I or II

1c.24 **Rationale for Using this Guideline Over Others:** These guidelines have been developed for healthcare personnel who insert intravascular catheters and for persons responsible for surveillance and control of infections in hospital, outpatient, and home healthcare settings. This report was prepared by a working group comprising members from professional organizations representing the disciplines of critical care medicine, infectious diseases, healthcare infection control, surgery, anesthesiology, interventional radiology, pulmonary medicine, pediatric medicine, and nursing. The working group was led by the Society of Critical Care Medicine (SCCM), in collaboration with the Infectious Diseases Society of America (IDSA), Society for Healthcare Epidemiology of America

(SHEA), Surgical Infection Society (SIS), American College of Chest Physicians (ACCP), American Thoracic Society (ATS), American Society of Critical Care Anesthesiologists (ASCCA), Association for Professionals in Infection Control and Epidemiology (APIC), Infusion Nurses Society (INS), Oncology Nursing Society (ONS), American Society for Parenteral and Enteral Nutrition (ASPEN), Society of Interventional Radiology (SIR), American Academy of Pediatrics (AAP), Pediatric Infectious Diseases Society (PIDS), and the Healthcare Infection Control Practices Advisory Committee (HICPAC) of the Centers for Disease Control and Prevention (CDC) and is intended to replace the Guideline for Prevention of Intravascular Catheter-Related Infections published in 2002. These guidelines are intended to provide evidence-based recommendations for preventing intravascular catheter-related infections.

Based on the NQF descriptions for rating the evidence, what was the <u>developer's assessment</u> of the quantity, quality, and consistency of the body of evidence?

1c.25 Quantity: High **1c.26** Quality: High**1c.27** Consistency**:** High

1. Evidence, Performance Gap, Priority – Importance to Measure and Report

Extent to which the specific measure focus is evidence-based, important to making significant gains in healthcare quality, and improving health outcomes for a specific high-priority (high-impact) aspect of healthcare where there is variation in or overall less-than-optimal performance. ***Measures must be judged to meet all subcriteria to pass this criterion and be evaluated against the remaining criteria.***

1a. Evidence to Support the Measure Focus – See attached Evidence Submission Form

1731_Evidence_MSF5.0_Data-635787040738097660.doc

1b. Performance Gap

Demonstration of quality problems and opportunity for improvement, i.e., data demonstrating:

- considerable variation, or overall less-than-optimal performance, in the quality of care across providers; and/or
- disparities in care across population groups.

1b.1. **Briefly explain the rationale for this measure** *(e.g., the benefits or improvements in quality envisioned by use of this measure)*

A health care-associated bloodstream infection in high-risk newborns remains a major patient safety concern. Effective preventive measures range from simple hand-washing protocols or closed medication delivery systems to more elaborate multidisciplinary quality improvement plans involving hand-washing, nutrition, skin care, respiratory care, vascular access, and diagnostic practices. Guidelines for the prevention of intravascular catheter-related infections are also available from the Centers for Disease Control and Prevention (CDC) to assist hospitals in establishing successful interventions to reduce the number of health care-associated bloodstream infections in newborns.

The measure will assist health care organizations (HCOs) to track evidence of a decrease in health care-associated bloodstream infections in newborns.

1b.2. **Provide performance scores on the measure as specified (current and over time) at the specified level of analysis.** *(This is required for endorsement maintenance. Include mean, std dev, min, max, interquartile range, scores by decile. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included). This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*

Health care-associated bloodstream infections continue to persist despite the fact that standardized guidelines have been developed for intravascular catheter care. The goal of eliminating health care-associated bloodstream infections has not been met for all hospitals reporting the data. The Perinatal Care (PC) core measures were added as a new core measure set in 2010 for hospitals to select in order to meet their ORYX performance measurement requirement for Joint Commission accreditation purposes. Approximately 158 hospitals reported the data with an average measure rate of 0.02% (n=45,248 patients). In January 2014, The Joint Commission required mandatory reporting of the PC measure set for all accredited hospitals with 1100 births or more annually. 1218 hospitals reported the data with an average rate of 3.2% (n=363,400 patients). The 2014 performance gap persists with improvement noted primarily in the median (0%), lower quartile (0%) and 10th percentile (0%) hospitals. It is important to note that a performance gap of 7.1% exists for the 90th percentile and 1.8% for the upper quartile of hospitals. The 2014 mean rate of 2.9% also remains above the target goal of 0%. The threshold for mandatory reporting was recently lowered to 300 births annually effective January 2016. The new reporting requirement will now capture approximately 80% of all accredited birthing hospitals. As a result, the rates may increase with the addition of approximately 821 more hospitals reporting data. Below is the specified level of analysis for PC-04 beginning with discharges April 1, 2010 through December 31, 2014.

2Q 2010: 45,248 denominator cases; 135 numerator cases; 158 hospitals; 0.02% national aggregate rate; 0.0036 mean of hospital rates; 0.01189 standard deviation; 0.08% 90th percentile rate; 0.02% 75th percentile rate/upper quartile; 0% 50th percentile rate/median rate; 0% 25th percentile rate/lower quartile; and 0% 10th percentile rate.

CY 2011: 6,490 denominator cases; 62 numerator cases; 109 hospitals; 0.09% national aggregate rate; 0.04755 mean of hospital rates; 0.19254 standard deviation; .3.8% 90th percentile rate; 0% 75th percentile rate/upper quartile; 0% 50th percentile rate/median rate; 0% 25th percentile rate/lower quartile; and 0% 10th percentile rate.

CY 2012: 2,570 denominator cases; 22 numerator cases; 89 hospitals; 0.08% national aggregate rate; 0.01566 mean of hospital rates; 0.10635 standard deviation; 2.2% 90th percentile rate; 0% 75th percentile rate/upper quartile; 0% 50th percentile rate/median rate; 0% 25th percentile rate/lower quartile; and 0% 10th percentile rate.

CY 2013: 4,861 denominator cases; 122 numerator cases; 129 hospitals; 2.5% national aggregate rate; 0.02693 mean of hospital rates; 0.10249 standard deviation; 6.2% 90th percentile rate; 2%% 75th percentile rate/upper quartile; 0% 50th percentile rate/median rate; 0% 25th percentile rate/lower quartile; and 0% 10th percentile rate.

CY 2014: 71,676 denominator cases; 2,297 numerator cases; 1218 hospitals; 3.2% national aggregate rate; 0.02981 mean of hospital rates; 0.09815 standard deviation; 7.1% 90th percentile rate; 1.8% 75th percentile rate/upper quartile; 0% 50th percentile rate/median rate; 0% 25th percentile rate/lower quartile; and 0% 10th percentile rate.

1b.3. **If no or limited performance data on the measure as specified is reported in 1b2, then provide a summary of data from the literature that indicates opportunity for improvement or overall less than optimal performance on the specific focus of measurement.**
Not Applicable

1b.4. **Provide disparities data from the measure as specified (current and over time) by population group, e.g., by race/ethnicity, gender, age, insurance status, socioeconomic status, and/or disability.** *(This is required for endorsement maintenance. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities include.) This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*
There is a great deal of literature supporting the standardization of aseptic care and educational interventions aimed at neonatal nurses to reduce the incidence of health care-associated bloodstream infections in newborns. There is no mention of disparities related to race or socioeconomic status regarding the incidence of health care associated bloodstream infections in newborns. Although the literature supports premature newborns with very low birth weight > 1500 g as the most vulnerable group of newborns susceptible to health care-associated bloodstream infections.

An updated 2015 literature search yielded no new information on disparities reported for health care-associated bloodstream infections in newborns. It is important to note that recent studies by Flett, et al. (2015) and Zachariah, et al. (2014) show increased compliance with standardized protocols proven to reduce the incidence of health care-associated bloodstream infections when state legal mandates for reporting of infection rates where in place.

1b.5. **If no or limited data on disparities from the measure as specified is reported in 1b4, then provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement. Include citations.**

- Adams-Chapman, I. & Stoll, B.J. (2002). Prevention of nosocomial infections in the neonatal intensive care unit. Current Opinion in Pediatrics.14 (2):157-64.
- Bloom, B.T., Craddock, A., Delmore, P.M., et al. (2003). Reducing acquired infections in the NICU: observing and implementing meaningful differences in process between high and low acquired infection rate centers. Journal of Perinatology. 23(6):489-92.
- Clark, R., Powers, R., White, R., Bloom, B., Sanchez, P., & Benjamin, D.K., Jr. (2004a). Prevention and treatment of nosocomial sepsis in the NICU. Journal of Perinatology. 4; 24(7):446-53.
- Clark, R., Powers, R., White, R., Bloom, B., Sanchez, P., & Benjamin, D.K., Jr. (2004b). Nosocomial infection in the NICU: a medical complication or unavoidable problem? Journal of Perinatology. 24(6):382-8.
- Gaynes, R.P., Edwards, J.R., Jarvis, W.R., Culver, D.H., Tolson, J.S., & Martone, W.J. (1996). Nosocomial infections among neonates in high-risk nurseries in the United States. National Nosocomial Infections Surveillance System. Pediatrics. 98(3 Pt 1):357-61.
- Payne, N.R., Carpenter, J.H., Badger, G.J., Horbar, J.D., & Rogowski, J. (2004). Marginal increase in cost and excess length of stay associated with nosocomial bloodstream infections in surviving very low birth weight infants. Pediatrics. 114(2):348-55.
- Sohn, A.H., Garrett, D.O., Sinkowitz-Cochran, R.L., et al. (2001). Prevalence of nosocomial infections in neonatal intensive care unit patients: Results from the first national point-prevalence survey. Journal of Pediatrics. 139(6):821-7.
- Stoll, B.J., Hansen, N., Fanaroff, A.A., et al. (2002). Late-onset sepsis in very low birth weight neonates: the experience of the NICHD Neonatal Research Network. Pediatrics. 110(2 Pt 1):285-91.
- Zachariah, P., Reagan, J., Furuya, E., Dick, A., Liu, H., Herzig, C., Pogorzelska-Maziarz, M., Stone, P. & Saiman, L. (2014). The association of state legal mandates for data submission of central line-associated bloodstream infections in neonatal intensive care units with process and outcome measures. Infect Control Hosp Epidemiol. 35(9):1133-9.

1c. High Priority (previously referred to as High Impact)

The measure addresses:

- a specific national health goal/priority identified by DHHS or the National Priorities Partnership convened by NQF; OR
- a demonstrated high-priority (high-impact) aspect of healthcare (e.g., affects large numbers of patients and/or has a substantial impact for a smaller population; leading cause of morbidity/mortality; high resource use (current and/or future); severity of illness; and severity of patient/societal consequences of poor quality).

1c.1. **Demonstrated high priority aspect of healthcare**
A leading cause of morbidity/mortality
1c.2. **If Other:**

1c.3. Provide epidemiologic or resource use data that demonstrates the measure addresses a high priority aspect of healthcare. List citations in 1c.4.

Health care-associated bacteremia is a significant problem for infants admitted into neonatal intensive care units (NICUs) and other hospital units. This is especially true for very low birth weight infants who are at high risk for these infections due to their immature immune systems and need for invasive monitoring and supportive care (Adams-Chapman & Stoll, 2002; Bloom et al., 2003; Clark et al., 2004a; Clark et al., 2004b; Gaynes et al., 1996; Payne et al., 2004; Sohn et al., 2001; Stoll et al., 2002). Mortality rates are high and infections result in increased length of stay as well as increased hospital costs and charges (Adams-Chapman & Stoll, 2002; Bloom et al.,2003; Clark et al., 2004b; Horbar et al., 2001; Kilbride et al., 2003; Sohn et al.,2001; Stoll et al., 2002).

The incidence of health care-associated bacteremia increases with decreasing birth weight. Other risk factors include central venous catheter use, prolonged time using parenteral nutrition, prolonged time on mechanical ventilation (Adams-Chapman & Stoll, 2002; Barton et al., 1999; Gaynes et al., 1996; Perlman et al., 2007.; Stoll et al., 2002). The most common causative organisms are coagulase-negative staphylococci, Staphylococcus aureus, enterococci, Enterobacter sp, and Escherichia coli (Adams-Chapman & Stoll, 2002; Clark et al., 2004b; Gaynes et al., 1996; Horbar et al., 2001; Payne et al., 2004; Sohn et al., 2001; Stoll et al., 2002). According to the National Nosocomial Infections Surveillance System data, the incidence of late-onset Methicillin-resistant Staphylococcus aureus (MRSA) infections in neonatal intensive care units increased substantially between1995 and 2004 (Lessa, et al., 2009).

Effective preventive measures range from simple hand-washing protocols or closed medication delivery systems to more elaborate multidisciplinary quality improvement plans involving hand-washing, nutrition, skin care, respiratory care, vascular access, and diagnostic practices. All of these interventions have been shown to substantially reduce infection rates, albeit in nonrandomized studies using historical or concurrent control units (Adams-Chapman & Stoll, 2002; Aly et al., 2005; Bloom et al., 2003; Clark et al., 2004a; Clark et al., 2004b; Horbar et al., 2001; Lam et al., 2004; Kilbride et al., 2003a; Kilbride et al., 2003b; Ng et al., 2004; Schelonka et al., 2006). A reduction from 24.6% to 16.4% was achieved with a multi-modality, multi-hospital intervention focusing on hand hygiene with an effective agent before and after every patient contact, eliminating hand jewelry and artificial nails, using maximal barrier precautions during central venous catheter insertion, decreasing the number of skin punctures, reducing the duration of intravenous lipid and deep line use, and improving the diagnosis of health care-associated infections. (Kilbride et al., 2003a; Kilbride et al., 2003b). In a review of educational interventions aimed at neonatal nurses, catheter-related bloodstream infection (CR-BSI) rates decreased by 40% in 9 studies and 21% in one study (Semelsberger, 2009).
Given the fragility and susceptibility of the patient population, a baseline level of health care-associated infections will be expected, even with good protocols in place. However, those centers that have prevention protocols, and are able to encourage health care workers to adhere to these protocols, will probably have success in reducing their rates of health care-associated bacteremia in their neonatal population. Indeed, several quasi-experimental studies have demonstrated that NICUs can lower their infection rates (based on positive blood cultures) from as high as 13.5 per 1,000 patient days to as low as 3.0 per 1,000 patient days(Adams-Chapman & Stoll, 2002; Aly et al., 2005; Bloom et al. ,2003; Clark et al., 2004a; Clark et al., 2004b; Horbar et al., 2001; Lam et al., 2004; Kilbride et al., 2003a; Kilbride et al., 2003b; Ng et al., 2004; Schelonka et al., 2006)

1c.4. Citations for data demonstrating high priority provided in 1a.3

• Adams-Chapman, I. & Stoll, B.J. (2002). Prevention of nosocomial infections in the neonatal intensive care unit. Current Opinion in Pediatrics.14 (2):157-64.
• Aly, H., Herson, V., Duncan, A., et al. (2005). Is bloodstream infection preventable among premature infants? A tale of two cities. Pediatrics. 115(6):1513-8.
• Barton, L., Hodgman, J.E., & Pavlova, Z. (1999). Causes of death in the extremely low birth weight infant. Pediatrics. 103(2):446-51.
• Bloom, B.T., Craddock, A., Delmore, P.M., et al. (2003). Reducing acquired infections in the NICU: observing and implementing meaningful differences in process between high and low acquired infection rate centers. Journal of Perinatology. 23(6):489-92.
• Clark, R., Powers, R., White, R., Bloom, B., Sanchez, P., & Benjamin, D.K., Jr. (2004a). Prevention and treatment of nosocomial sepsis in the NICU. Journal of Perinatology. 4; 24(7):446-53.
• Clark, R., Powers, R., White, R., Bloom, B., Sanchez, P., & Benjamin, D.K., Jr. (2004b). Nosocomial infection in the NICU: a medical complication or unavoidable problem? Journal of Perinatology. 24(6):382-8.
• Gaynes, R.P., Edwards, J.R., Jarvis, W.R., Culver, D.H., Tolson, J.S., & Martone, W.J. (1996). Nosocomial infections among neonates in high-risk nurseries in the United States. National Nosocomial Infections Surveillance System. Pediatrics. 98(3 Pt 1):357-61.
• Horbar, J.D., Rogowski, J., Plsek, P.E., et al. (2001). Collaborative quality improvement for neonatal intensive care. NIC/Q Project Investigators of the Vermont Oxford Network. Pediatrics. 107(1):14-22.
• Kilbride, H.W., Wirtschafter, D.D., Powers, R.J., & Sheehan, M.B. (2003a). Implementation of evidence-based potentially better practices to decrease nosocomial infections. Pediatrics. 111(4 Pt 2):e519-33.
• Kilbride, H.W., Powers, R., Wirtschafter, D.D., et al. (2003b). Evaluation and development of potentially better practices to prevent neonatal nosocomial bacteremia. Pediatrics. 111(4 Pt 2):e504-18.

• Lam, B.C., Lee, J., & Lau, Y.L. (2004). Hand Hygiene Practices in a Neonatal Intensive Care Unit: A Multimodal Intervention and Impact on Nosocomial Infection. Pediatrics.114 (5):e565.
• Lessa, F.C., Edwards, J.R., Fridkin, S.K., Tenover, F.C. & Gorwitz, R.J. (2009). Trends in incidence of late-onset Methicillin-resistant Staphylococcus aureus infectin in neonatal intensive care units: data from the National Nosocomial Infections Surveillance System, 1995-2004. Pediatric Infect. Dis, J. 28 (7).577-81.
• Neill, S., Haithcock, S., Smith, P., Goldberg, R., Bidegain, M., Tanaka, D., Carriker, C. & Ericson, J. (2015). Sustained reduction in bloodstream infections in infants at a large tertiary care neonatal intensive care unit. Adv Neonatal Care. [Epub ahead of print]
• Ng, P.C., Wong, H.L., Lyon, D.J., et al. (2004). Combined use of alcohol hand rub and gloves reduces the incidence of late onset infection in very low birthweight infants. Archives of Disease in Childhood Fetal & Neonatal Edition. 89(4):F336-40.
• Payne, N.R., Carpenter, J.H., Badger, G.J., Horbar, J.D., & Rogowski, J. (2004). Marginal increase in cost and excess length of stay associated with nosocomial bloodstream infections in surviving very low birth weight infants. Pediatrics. 114(2):348-55.
• Perlman, S.E., Saiman, L., Larson, E.L. (2007). Risk factors for late-onset health care- associated bloodstream infections in patients in neonatal intensive care units. Am J. Infect Control Apr;35 (3):177-82.
• Schelonka, R.L., Scruggs, S., Nichols, K., Dimmitt, R.A., & Carlo, W.A. (2006). Sustained reductions in neonatal nosocomial infection rates following a comprehensive infection control intervention. Journal of Perinatology. 26(3):176-9.
• Semelsberger, C.F. (2009). Educational interventions to reduce the rate of catheter-related bloodstream infections in the NICU: a review of the research literature. Neonatal Netw. 28(6) 391-5.
• Shepherd, E., Kelly, T., Vinsel, J., Cunningham, D., Keels, E., Beauseau, W. & McClead, R. Jr. (2015). Significant reduction of central-line associated bloodstream infections in a network of diverse neonatal nurseries. J Pediatr. 167(1):41-6.e1-3.
• Sohn, A.H., Garrett, D.O., Sinkowitz-Cochran, R.L., et al. (2001). Prevalence of nosocomial infections in neonatal intensive care unit patients: Results from the first national point-prevalence survey. Journal of Pediatrics. 139(6):821-7.
• Stoll, B.J., Hansen, N., Fanaroff, A.A., et al. (2002). Late-onset sepsis in very low birth weight neonates: the experience of the NICHD Neonatal Research Network. Pediatrics. 110(2 Pt 1):285-91.

1c.5. **If a PRO-PM (e.g. HRQoL/functional status, symptom/burden, experience with care, health-related behaviors), provide evidence that the target population values the measured PRO and finds it meaningful.** *(Describe how and from whom their input was obtained.)*
Not Applicable

2. Reliability and Validity—Scientific Acceptability of Measure Properties

Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. ***Measures must be judged to meet the subcriteria for both reliability and validity to pass this criterion and be evaluated against the remaining criteria.***

2a.1. Specifications The measure is well defined and precisely specified so it can be implemented consistently within and across organizations and allows for comparability. eMeasures should be specified in the Health Quality Measures Format (HQMF) and the Quality Data Model (QDM).

De.5. **Subject/Topic Area** *(check all the areas that apply):*
Perinatal and Reproductive Health

De.6. **Cross Cutting Areas** *(check all the areas that apply):*
Safety : Healthcare Associated Infections

S.1. **Measure-specific Web Page** *(Provide a URL link to a web page specific for this measure that contains current detailed specifications including code lists, risk model details, and supplemental materials. Do not enter a URL linking to a home page or to general information.)*
http://manual.jointcommission.org/releases/TJC2015B2/

S.2a. **If this is an eMeasure**, HQMF specifications must be attached. Attach the zipped output from the eMeasure authoring tool (MAT) - if the MAT was not used, contact staff. (Use the specification fields in this online form for the plain-language description of the specifications)
This is not an eMeasure **Attachment:**

S.2b. **Data Dictionary, Code Table, or Value Sets** *(and risk model codes and coefficients when applicable) must be attached. (Excel or csv file in the suggested format preferred - if not, contact staff)*
Attachment **Attachment:** PC04_ICD_Code_Tables.xlsx

S.3. **For endorsement maintenance**, please briefly describe any changes to the measure specifications since last endorsement date and explain the reasons.

- All ICD-9-CM diagnosis codes and ICD-9-CM procedure codes were converted to ICD-10-CM diagnosis and ICD-10-PCS procedure codes throughout the measure specifications. The Department of Health and Human Services (HHS) mandated that entities covered by the Health Insurance Portability and Accountability Act (HIPAA) must all transition to a new set of codes for electronic health care transactions on October 1, 2015.
- The numerator included population now requires a check to confirm that the bloodstream infection was health care-associated after the first 48 hours when infection codes are present on Table 11.10 or 11.10.1 with a new data element Bloodstream Infection Confirmed, since infection codes are also applied for infections resulting from other newborn medical conditions that are not health care–associated, i.e., necrotizing enterocolitis, pneumonia, urosepsis, etc.
- The initial patient population now identifies eligible patients with ICD-10-CM-Principal Diagnosis Code for newborn born at hospital and data elements Point of Origin for Admission or Visit and Admission Type were removed, since updates to the latter data elements are no longer available in the public domain.
- Table 11.17 for birth weight was removed from the denominator included population, since there are no equivalent birth weight codes after the conversion to ICD-10-CM diagnosis codes.
- The length of stay > 120 days was removed from the denominator excluded population, since many low birth weight newborns have longer lengths of stay and were being excluded from the measure.
- Vital Records reports, delivery logs and clinical information systems were added as acceptable data sources to help hospitals identify birth weights in order to reduce the burden of data abstraction when cases are not routinely coded with diagnosis codes for birth weights.
- A new denominator data element Bloodstream Infection Present on Admission was added in order to exclude newborns with ICD-10-CM diagnosis codes for septicemias and bacteremias from the denominator who had bloodstream infections present on admission.

S.4. **Numerator Statement** *(Brief, narrative description of the measure focus or what is being measured about the target population, i.e., cases from the target population with the target process, condition, event, or outcome)*
IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.

The outcome being measured is: Newborns with septicemia or bacteremia with ICD-10-CM Other Diagnosis Codes for newborn septicemia or bacteremia as defined in Appendix A, Table 11.10 with a Bloodstream Infection Confirmed OR ICD-10-CM Other Diagnosis Codes for sepsis as defined in Appendix A, Table 11.10.1 with a Bloodstream Infection Confirmed available at: http://manual.jointcommission.org/releases/TJC2015B2/

The only national hospital quality measure currently requiring patient-level risk adjustment is the Health Care-Associated Bloodstream Infections in Newborns (PC-04) outcome measure in the perinatal care measure set.

S.5. **Time Period for Data** *(What is the time period in which data will be aggregated for the measure, e.g., 12 mo, 3 years, look back to August for flu vaccination? Note if there are different time periods for the numerator and denominator.)*
Episode of care

S.6. **Numerator Details** *(All information required to identify and calculate the cases from the target population with the target process, condition, event, or outcome such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
IF an OUTCOME MEASURE, describe how the observed outcome is identified/counted. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
Two data elements are used for the observed outcome and to calculate the numerator:

1. Bloodstream Infection Confirmed- Confirmation that a health care-associated bloodstream infection occurred after the first 48 hours after admission.
2. ICD-10-CM Other Diagnosis Codes- The International Classification of Diseases, Tenth Revision, Clinical Modification codes associated with the secondary diagnoses for this hospitalization.

Cases are eligible for the numerator population with ICD-10-CM Other Diagnosis Code for newborn septicemia or bacteremia with the presence of a health care-associated bloodstream infection confirmed OR an ICD-10-CM Other Diagnosis Codes for sepsis with the presence of a health care-associated bloodstream infection confirmed.
Updates available at: https://manual.jointcommission.org/releases/TJC2015B2/.

S.7. **Denominator Statement** *(Brief, narrative description of the target population being measured)*
The outcome target population being measured is: Liveborn newborns with ICD-10-CM Other Diagnosis Codes for birth weight between 500 and 1499g as defined in Appendix A, Table 11.12, 11.13 or 11.14 OR Birth Weight between 500 and 1499g OR ICD-10-CM Other Diagnosis Codes for birth weight = > 1500g as defined in Appendix A, Table 11.15 or 11.16 OR Birth Weight = > 1500g who experienced one or more of the following:
o Experienced death
o ICD-10-PCS Principal Procedure Code or ICD-10-PCS Other Procedure Codes for major surgery as defined in Appendix A, Table 11.18
o ICD-10-PCS Principal Procedure Code or ICD-10-PCS Other Procedure Codes for mechanical ventilation as defined in Appendix A, Table 11.19
o Transferred in from another acute care hospital or health care setting within 2 days of birth.

S.8. **Target Population Category** *(Check all the populations for which the measure is specified and tested if any):*
Populations at Risk

S.9. **Denominator Details** *(All information required to identify and calculate the target population/denominator such as definitions, specific data collection items/responses , code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
Eleven data elements are used to identify the target population and to calculate the denominator:

1. Admission Date – The month, day and year of admission to acute inpatient care.
2. Birth Weight- The weight (in grams) of a newborn at the time of delivery.
3. Birthdate - The month, day and year the patient was born.
4. Bloodstream Infection Present on Admission- Documentation in the medical record that the patient had a bloodstream infection present on admission. This includes both patients with positive blood cultures or inconclusive blood cultures when the patient is suspected of having a bloodstream infection or septicemia and is being treated for the condition. Allowable values: Yes or No/UTD

5. Clinical Trial - Documentation that during this hospital stay the patient was enrolled in a clinical trial in which patients who are newborns were being studied. Allowable values: Yes or No/UTD
6. Discharge Date – The month day and year the patient was discharged from acute care, left against medical advice or expired during the stay.
7. Discharge Disposition - The place or setting to which the patient was discharged.
8. ICD-10-CM Other Diagnosis Codes - The International Classification of Diseases, Tenth Revision, Clinical Modification codes associated with the secondary diagnoses for this hospitalization.
9. ICD-10-PCS Other Procedure Codes - The International Classification of Diseases, Tenth Revision, Procedure Coding System code that identifies significant procedures performed other than the principal procedure during this hospitalization.
10. ICD-10-CM Principal Diagnosis Code - The International Classification of Diseases, Tenth Revision, Clinical Modification code associated with the diagnosis established after study to be chiefly responsible for occasioning the admission of the patient for this hospitalization.
11. ICD-10-PCS Principal Procedure Code - The International Classification of Diseases, Tenth Revision, Procedure Coding System code that identifies the principal procedure performed during this hospitalization. The principal procedure is the procedure performed for definitive treatment rather than diagnostic or exploratory purposes, or which is necessary to take care of a complication.
Updates available at: https://manual.jointcommission.org/releases/TJC2015B2/.

S.10. Denominator Exclusions *(Brief narrative description of exclusions from the target population)*
• ICD-10-CM Principal Diagnosis Code for septicemias or bacteremias as defined in Appendix A, Table 11.10.2
• ICD-10-CM Other Diagnosis Codes for septicemias or bacteremias as defined in Appendix A, Table 11.10.2 or ICD-10-CM Principal or Other Diagnosis Codes for newborn septicemia or bacteremia as defined in Appendix A, Table 11.10 with a Bloodstream Infection Present on Admission
• ICD-10-CM Other Diagnosis Codes for birth weight < 500g as defined in Appendix A, Table 11.20 OR Birth Weight < 500g
• Length of Stay < 2 days
• Enrolled in clinical trials

S.11. Denominator Exclusion Details *(All information required to identify and calculate exclusions from the denominator such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
• Patients with ICD-10-CM Principal Diagnosis Code for septicemias or bacteremias are excluded.
• Patients with ICD-10-CM Other Diagnosis Codes for septicemias or bacteremias with a Bloodstream Infection Present on Admission are excluded.
• Patients with ICD-10-CM Principal or Other Diagnosis Codes for newborn septicemia or bacteremia with a Bloodstream Infection Present on Admission are excluded.
• Patients with ICD-10-CM Other Diagnosis Codes for birth weight <500 grams OR a birth weight <500 grams are excluded.
• Length of stay (LOS) in days is equal to the Discharge Date minus the Admission Date. If the LOS is less than 2 days, the patient is excluded.
• Patients are excluded if "Yes" is selected for Clinical Trial.

S.12. Stratification Details/Variables *(All information required to stratify the measure results including the stratification variables, definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format with at S.2b)*
Not applicable, the measure is not stratified.

S.13. Risk Adjustment Type (Select type. Provide specifications for risk stratification in S.12 and for statistical model in S.14-15)
Statistical risk model
If other:

S.14. Identify the statistical risk model method and variables *(Name the statistical method - e.g., logistic regression and list all the risk factor variables. Note - risk model development and testing should be addressed with measure testing under Scientific Acceptability)*
Logistic regression
Model Risk Factors Considered:
Intercept Intercept
Birth Weight 1250g to 2499g
Birth Weight 1000 to 1249g

Birth Weight 500 to 749g
Birth Weight 750 to 750g
Modified DRG Newborn Transfers Out or Died
Congenital Anomaly Gastrointestinal Anomaly
Congenital Anomaly Cardiovascular Anomaly
Congenital Anomaly Other Anomaly
Out-born Birth Newborns Transfers In

S.15. **Detailed risk model specifications** *(must be in attached data dictionary/code list Excel or csv file. Also indicate if available at measure-specific URL identified in S.1.)*
Note: Risk model details (including coefficients, equations, codes with descriptors, definitions), should be provided on a separate worksheet in the suggested format in the Excel or csv file with data dictionary/code lists at S.2b.
Available in attached Excel or csv file at S.2b

S.15a. **Detailed risk model specifications** (*if not provided in excel or csv file at S.2b*)

S.16. **Type of score:**
Rate/proportion
If other:

S.17. **Interpretation of Score** *(Classifies interpretation of score according to whether better quality is associated with a higher score, a lower score, a score falling within a defined interval, or a passing score)*
Better quality = Lower score

S.18. **Calculation Algorithm/Measure Logic** *(Describe the calculation of the measure score as an ordered sequence of steps including identifying the target population; exclusions; cases meeting the target process, condition, event, or outcome; aggregating data; risk adjustment; etc.)*
1. Start processing. Run cases that are included in the PC-Newborn Initial Patient Newborns with BSI and pass the edits defined in the Transmission Data Processing Flow: Clinical through this measure.

2. Calculate Length of Stay. Length of Stay, in days, is equal to the Discharge Date minus the Admission Date.

3. Check Length of Stay
a. If Length of Stay is less than 2 days, the case will proceed to a Measure Category Assignment of B and will not be in the measure population. Stop processing.
b. If Length of Stay is greater than or equal to 2 days, continue processing and proceed to Clinical Trial.

4. Check Clinical Trial
a. If Clinical Trial is missing, the case will proceed to a Measure Category Assignment of X and will be rejected. Stop processing.
b. If Clinical Trial equals Yes, the case will proceed to a Measure Category Assignment of B and will not be in the measure population. Stop processing.
c. If Clinical Trial equals No, continue processing and proceed to ICD-10-CM Principal or Other Diagnosis Codes.

5. Check ICD-10-CM Principal or Other Diagnosis Codes
a. If none of the ICD-10-CM Principal or Other Diagnosis Codes is on Table 11.10, continue processing and proceed to ICD-10-CM Other Diagnosis Codes
1. If all of the ICD-10-CM Other Diagnosis Codes are missing or none of the ICD-10-CM Other Diagnosis Codes is on Table 11.10.2, continue processing and proceed to recheck ICD-10-CM Other Diagnosis Codes (Step 7).
2. If at least one of the ICD-10-CM Other Diagnosis Codes is on Table 11.10.2, continue processing and proceed to Bloodstream Infection Present on Admission.
b. If at least one of the ICD-10-CM Principal or Other Diagnosis Codes is on Table 11.10, continue processing and proceed to Bloodstream Infection Present on Admission.

6. Check Bloodstream Infection Present on Admission
a. If Bloodstream Infection Present on Admission is missing, the case will proceed to a Measure Category Assignment of X and will be rejected. Stop processing.

b. If Bloodstream Infection Present on Admission equals Yes, the case will proceed to a Measure Category Assignment of B and will not be in the measure population. Stop processing.
c. If Bloodstream Infection Present on Admission equals No, continue processing and proceed to check ICD-10-CM Other Diagnosis Codes.

7. Check ICD-10-CM Other Diagnosis Codes
a. If at least one of the ICD-10-CM Other Diagnosis Codes is on Table 11.12, 11.13, 11.14, continue processing and proceed to recheck ICD-10-CM Other Diagnosis Codes (Step 13).
b. If all of the ICD-10-CM Other Diagnosis Codes are missing, continue processing and proceed to Birth Weight.
c. If none of the ICD-10-CM Other Diagnosis Codes is on Table 11.12, 11.13, 11.14, continue processing and proceed to recheck ICD-10-CM Other Diagnosis Codes (Step 8).

8. Recheck ICD-10-CM Other Diagnosis Codes
a. If at least one of the ICD-10-CM Other Diagnosis Codes on table 11.15, 11.16, continue processing and proceed to ICD-10-CM Principal or Other Procedure Codes.
b. If none of the ICD-10-CM Other Diagnosis Codes on table 11.15, 11.16, continue processing and proceed to Birth Weight.

9. Check Birth Weight
a. If Birth Weight is missing, the case will proceed to a Measure Category Assignment of X and will be rejected. Stop processing.
b. If Birth Weight equals a Non Unable to Determine Value, the case will proceed to a Measure Category Assignment of E and will be in the Numerator Population. Stop processing.
c. If Birth Weight is less than 500, the case will proceed to a Measure Category Assignment of B and will not be in the measure population. Stop processing.
d. If Birth Weight is between 500 and 1499, continue processing and proceed to recheck ICD-10-CM Other Diagnosis Codes (Step13).
e. If Birth Weight is greater than or equal to 1500, continue processing and proceed to ICD-10-PCS Principal or Other Procedure Codes.

10. Check ICD-10-PCS Principal or Other Procedure Codes
a. If at least one of the ICD-10-PCS Principal or Other Procedure Codes is on table 11.18 or 11.19, continue processing and proceed to recheck ICD-10-PCS Other Diagnosis Codes (Step 13).
b. If all of the ICD-10-PCS Principal or Other Procedure Codes are missing or none of the ICD-10-PCS Principal or Other Procedure Codes is on table 11.18 or 11.19, continue processing and proceed to ICD-10-CM Principal Diagnosis Code.

11. Check ICD-10-CM Principal Diagnosis Code
a. If ICD-10-CM Principal Diagnosis Code is not on table 11.10.3, continue processing and proceed to recheck ICD-10-CM Other Diagnosis Codes (Step 13).
b. If ICD-10-CM Principal Diagnosis Code is on table 11.10.3, continue processing and proceed to Discharge Disposition.

12. Check Discharge Disposition
a. If Discharge Disposition is missing, the case will proceed to a Measure Category Assignment of X and will be rejected. Stop processing.
b. If Discharge Disposition equals 1, 2, 3, 4, 5, 7, 8, the case will proceed to a Measure Category Assignment of B and will not be in the measure population. Stop processing.
c. If Discharge Disposition equals 6, continue processing and proceed to recheck ICD-10-CM Other Diagnosis Codes (Step13).

13. Recheck ICD-10-CM Other Diagnosis Codes
a. If at least one of the ICD-10-CM Other Diagnosis Codes is on table 11.10, continue processing and proceed to Bloodstream Infection Confirmed.
b. If all of the ICD-10-CM Other Diagnosis Codes are missing or none of the ICD-10-CM Other Diagnosis Codes is on table 11.10, continue processing and proceed to recheck ICD-10-CM Other Diagnosis Codes (Step14).

14. Recheck ICD-10-CM Other Diagnosis Codes
a. If at least one of the ICD-10-CM Other Diagnosis Codes is on table 11.10.1, continue processing and proceed to Bloodstream Infection Confirmed.
b. If all of the ICD-10-CM Other Diagnosis Codes are missing or none of the ICD-10-CM Other Diagnosis Codes is on table 11.10.1, the case will proceed to a Measure Category Assignment of D and will be in the Measure Population. Stop processing.

15. Check Bloodstream Infection Confirmed
a. If Bloodstream Infection Confirmed is missing, the case will proceed to a Measure Category Assignment of X and will be rejected. Stop processing.
b. If Bloodstream Infection Confirmed equals Yes, the case will proceed to a Measure Category Assignment of E and will be in the Numerator Population. Stop processing.
c. If Bloodstream Infection Confirmed equals No, the case will proceed to a Measure Category Assignment of D and will be in the Measure Population. Stop processing.

Calculation of adjusted outcome:

Step 1 -- Identify the measure population through Measure Category Assignments.
Risk adjusted rate-based measure: Identify the numerator (Measure Category Assignment = E) and the denominator (Measure Category Assignment = D) cases using the information provided in the Measure Information Form (MIF). Risk adjusted continuous variable measure: Identify the number of cases in the measure population (Measure Category Assignment = D). At this time, there are no risk adjusted continuous outcome measures in any of the national hospital quality measure sets.
Note: Do not calculate a Predicted Value for a case if it is rejected by front-end edits or is
rejected because one or more measures in the measure set evaluates to a Measure Category Assignment = X.

Step 2 -- Create risk factors for the measure.
Using the Risk Model Information File provided by the Joint Commission, identify all applicable EOC record data elements and the associated risk factor values for each of the EOC records identified instep 1. Risk factors include patient demographic and/or clinical factors, which can influence outcomes of care. Some examples of risk factors include age, sex, and comorbidities – such as diabetes or a history of hypertension. As an example, Figure 1 lists the data elements required for risk adjustment of generic measure 'ABC'. Using the data for measure 'ABC', the performance measurement system must identify the risk factors at the EOC record-level, and create data subsets for each participating hospital.

S.19. **Calculation Algorithm/Measure Logic Diagram URL or Attachment** *(You also may provide a diagram of the Calculation Algorithm/Measure Logic described above at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
Available at measure-specific web page URL identified in S.1

S.20. **Sampling** *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*
IF a PRO-PM, identify whether (and how) proxy responses are allowed.
Sampling is not allowed for this measure.

S.21. **Survey/Patient-reported data** *(If measure is based on a survey, provide instructions for conducting the survey and guidance on minimum response rate.)*
IF a PRO-PM, specify calculation of response rates to be reported with performance measure results.
Not Applicable

S.22. **Missing data** (specify how missing data are handled, e.g., imputation, delete case.)
Required for Composites and PRO-PMs.
Any file with missing data will result in a measure category assignment of X and rejection of the file from the warehouse, unless the data are not used to process the measure. If the data are used to process the measure and have been reported by the abstractor as No/UTD, the case will result in a measure category assignment of E (i.e., failed measure, not rejected).

S.23. **Data Source** *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*
If other, please describe in S.24.
Paper Medical Records

S.24. **Data Source or Collection Instrument** *(Identify the specific data source/data collection instrument e.g. name of database, clinical registry, collection instrument, etc.)*
IF a PRO-PM, identify the specific PROM(s); and standard methods, modes, and languages of administration.
Each data element in the data dictionary includes suggested data sources. The data are collected using contracted Performance Measurement Systems (vendors) that develop data collection tools based on the measure specifications. The tools are verified and tested by Joint Commission staff to confirm the accuracy and conformance of the data collection tool with the measure specifications. The vendor may not offer the measure set to hospitals until verification has been passed.

S.25. **Data Source or Collection Instrument** *(available at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
No data collection instrument provided

S.26. **Level of Analysis** *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*
Facility, Population : National

S.27. **Care Setting** *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*
Hospital/Acute Care Facility
If other:

S.28. **COMPOSITE Performance Measure** - Additional Specifications *(Use this section as needed for aggregation and weighting rules, or calculation of individual performance measures if not individually endorsed.)*
Not Applicable

2a. Reliability – See attached Measure Testing Submission Form
2b. Validity – See attached Measure Testing Submission Form
PC-04_1731_MeasureTesting_MSF5.0_Data-635787040739345660.doc

NATIONAL QUALITY FORUM

Measure missing data in MSF 6.5 from MSF 5.0

NQF #: 1731 **NQF Project:** Perinatal and Reproductive Health Project

2. RELIABILITY & VALIDITY - SCIENTIFIC ACCEPTABILITY OF MEASURE PROPERTIES

Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. **(evaluation criteria)**
Measure testing must demonstrate adequate reliability and validity in order to be recommended for endorsement. Testing may be conducted for data elements and/or the computed measure score. Testing information and results should be entered in the appropriate field. Supplemental materials may be referenced or attached in item 2.1. See guidance on measure testing.

2a2. Reliability Testing. (*Reliability testing was conducted with appropriate method, scope, and adequate demonstration of reliability.*)

2a2.1 Data/Sample *(Description of the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)*:
The PC measure set has been in national use since the 2nd quarter of 2010. It is a requirement of participation in the ORYX initiative that data on all measures in the set are collected. (ORYX is the term used by The Joint Commission to describe the component of the hospital accreditation program which requires data collection and reporting on standardized national performance measures.)
Demographics of organizations collecting and reporting data on these measures i are as follows:
163 health care organizations representing various types, locations and sizes:
10 For Profit, 91 Not for Profit, 46 Military Facilities, 9 County, 2 State, 5 Other
15 >=500 beds; 29 250-499 beds; 50 100-249 beds; 69 <100 beds
Located in: AE, AK, AL, AP, AR, AZ, CA, DO, DC, FL, GA, HI, ID, IL, IN, KS, KY, LA, MA, MD, MI, MN, MO, MS, MT, NC, NE, NV, NY, OH,OK, PA, PR, RI, SC, TN, TX, VA, WA, WI, WV
26 performance measurement systems

2a2.2 Analytic Method *(Describe method of reliability testing & rationale)*:
This measure was adapted from NQF-endorsed measure 0478 Nosocomial Blood Stream Infection in Neonates. As such, reliability was addressed during the original endorsement. The Joint Commission will be conducting further reliability studies on the entire PC measure set beginning in late 2011.

Currently, hospitals are supported in their data collection and reporting efforts by 26 contracted performance measurement system (PMS) vendors. It is a contractual requirement of Joint Commission listed vendors that the quality and reliability of data submitted to them by contracted health care organizations must be monitored on a quarterly basis. In addition, The Joint Commission analyzes these data by running 17 quality tests on the data submitted into ORYX. (ORYX is the term used by The Joint Commission to describe the component of the hospital accreditation program which requires data collection and reporting on standardized national performance measures). The following is a list of the major tests done on the submitted ORYX data, taken from the 2011 ORYX Performance Measurement System Requirements manual.

- Transmission of complete data
- Usage of individual core measure data received: To understand if the HCO provides the relevant service to treat the relevant population
- Investigation of aberrant data points
- Verification of patient population and sample size
- Identification of missing data elements
- Validation of the accuracy of target outliers
- Data integrity
- Data corrections

Data Element Agreement Rate:

Inter-rater reliability testing methodology utilized by contracted performance measure system vendors as outlined in the contract is as follows:

- All clinical data elements and all editable demographic elements are scored.
- All measure data are reabstracted with originally abstracted data having been blinded so that the reabstraction is not biased.
- Reabstracted data are compared with originally abstracted data on a data element by data element basis. A data element agreement rate is calculated. Clinical and demographic data are scored separately, and an overall agreement rate is computed.

2a2.3 Testing Results *(Reliability statistics, assessment of adequacy in the context of norms for the test conducted)***:**
Data element agreement rates were reported to The Joint Commission for 1Q11 This reflects the findings of 106 hospitals, comprising 26,302 records (100% sample). The following table delineates calculated agreement rates for individual data elements that are used to compute measure rates for PC-04.

Data Elements with a Mismatch - Newborn	total n	total d	rate
Admission Date	661	662	99.85%
Admission Type	661	662	99.85%
Birth Weight	623	662	94.11%
Point of Origin for Admission or Visit	671	672	99.85%

These agreement rates are considered to be well within acceptable levels.

2b. VALIDITY. Validity, Testing, <u>including all Threats to Validity</u>: H☐ M☐ L☐ I ☐

2b1.1 Describe how the measure specifications *(measure focus, target population, and exclusions)* **are consistent with the evidence cited in support of the measure focus (***criterion 1c)* **and identify any differences from the evidence:**
This measure focuses on the rate of health care-associated bloodstream infections in vulnerable newborns. The literature supports the focus on very low birth weight newborns with immature immune systems and newborns requiring mechanical ventilation and invasive procedures. Accordingly, this measure excludes newborns born with an infection at both the reporting and transferring hospital and those with extreme prematurity with a birth weight less than 500 gms. Also excluded from the measure are patients with a length of stay greater than 120 days, and those enrolled in a clinical trial. These exclusions are not addressed in the literature, but are included for this measure in order to harmonize with other CMS/Joint Commission aligned measures. In addition, those with a length of stay less than 2 days are excluded since these patients will typically be transferred to a higher level of care who will track these patients. The use of this measure reflects The Joint Commission's original 2006, and renewed 2011, National Patient Safety Goal for Hospitals, Requirement 07.03.01 to "Implement evidence-based practices to prevent health care–associated infections due to multidrug-resistant organisms in acute care hospitals" and the 2009 and renewed Requirement 07.04.01 to "Implement evidence-based practices to prevent central line–associated bloodstream infections".

2b2. Validity Testing. (*Validity testing was conducted with appropriate method, scope, and adequate demonstration of validity.)*

2b2.1 Data/Sample *(Description of the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)***:**
The PC measure set has been in national use since the 2nd quarter of 2010. Demographics of organizations collecting and reporting data on these measures s are as follows:
163 health care organizations representing various types, locations and sizes:
10 For Profit, 91 Not for Profit, 46 Military Facilities, 9 County, 2 State, 5 Other
15 >=500 beds; 29 250-499 beds; 50 100-249 beds; 69 <100 beds

Located in: AE, AK, AL, AP, AR, AZ, CA, DO, DC, FL, GA, HI, ID, IL, IN, KS, KY, LA, MA, MD, MI, MN, MO, MS, MT, NC, NE, NV, NY, OH,OK, PA, PR, RI, SC, TN, TX, VA, WA, WI, WV

2b2.2 Analytic Method *(Describe method of validity testing and rationale; if face validity, describe systematic assessment)***:**
Since the measure has been in national use, continued face validity of the measure has been determined through analysis of feedback from measure users. The Joint Commission provides a web-based application with which measure users can provide feedback regarding appropriateness of measure specifications, request clarification of specifications, and/or provide other comments pertinent to the measure. This feedback is systematically continually reviewed in order to identify trends and to identify areas of the measure specifications that require clarification or revision. Additionally, Joint Commission staff continually monitors the national literature and environment in order to assess continued validity of this measure.

In addition, The Joint Commission will begin reliability site visits this year. During the site visits, Joint Commission staff will conduct

focus group interviews with hospital staff working with the PC measures to obtain feedback regarding the validity of the measures and suggestions for further refinement of the specifications.

2b2.3 **Testing Results** *(Statistical results, assessment of adequacy in the context of norms for the test conducted; if face validity, describe results of systematic assessment)***:**
Analysis of feedback obtained via our automated feedback system reveals slightly more than 20 submissions regarding specifications for this measure since its implementation in 2010. Predominant themes of these submissions involved questions regarding clarification of the data elements Point of Origin for Admission or Visit and Admission with respect to classification of births outside of the hospital. Additional notes for abstractors were added to the data elements for clarification. Other notes for abstractors were added to the data element admission date to clarify the date of delivery is used as the admission date and not the date of the order written to admit. In addition, the denominator excluded population and algorithm were revised to capture newborns born with infections born both inside and outside of the reporting hospital with additional ICD-9-CM diagnosis code tables.

POTENTIAL THREATS TO VALIDITY. (*All potential threats to validity were appropriately tested with adequate results.*)

2b3. Measure Exclusions. (*Exclusions were supported by the clinical evidence in 1c or appropriately tested with results demonstrating the need to specify them.*)

2b3.1 **Data/Sample for analysis of exclusions** *(Description of the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)***:**
The PC measure set has been in national use since the 2nd quarter of 2010. Demographics of organizations collecting and reporting data on these measures are as follows:
163 health care organizations representing various types, locations and sizes:
10 For Profit, 91 Not for Profit, 46 Military Facilities, 9 County, 2 State, 5 Other
15 >=500 beds; 29 250-499 beds; 50 100-249 beds; 69 <100 beds

Located in: AE, AK, AL, AP, AR, AZ, CA, DO, DC, FL, GA, HI, ID, IL, IN, KS, KY, LA, MA, MD, MI, MN, MO, MS, MT, NC, NE, NV, NY, OH,OK, PA, PR, RI, SC, TN, TX, VA, WA, WI, WV
26 performance measurement systems

2b3.2 **Analytic Method** (*Describe type of analysis and rationale for examining exclusions, including exclusion related to patient preference)***:**
Measure exclusions that were not derived directly from the evidence are presented below. Please note that these are population exclusions that are necessary to ensure consistency in all measures in this 5 measure set.

These exclusions were analyzed for frequency of occurrence. An issue that is of great concern to users of this measure is that due to the presence of exceptions to the measure, attainment of a 0% measure rate is not possible. Because of the role of this measure in the current Joint Commission accreditation process, this is especially troubling to measure users. This concern is the basis for a number of the non-evidence-based exclusions to these measures. above. The following measure exclusions that were not derived directly from the evidence are as follows:

1. Patients with LOS <120 days
2. Patients with LOS >2 days
3. Patients enrolled in clinical trials

2b3.3 **Results** *(Provide statistical results for analysis of exclusions, e.g., frequency, variability, sensitivity analyses):*
N=356,671
1. Patients who have a length of stay (LOS) less than two days and greater than 120 days =3.64%
2. Patients enrolled in clinical trials =0.02%

2b4. Risk Adjustment Strategy. (*For outcome measures, adjustment for differences in case mix (severity) across measured entities was appropriately tested with adequate results.*)

2b4.1 **Data/Sample** *(Description of the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)***:**
Data source for the risk model was the 2007 State Inpatient Data, Healthcare Cost and Utilization Project, Agency for Healthcare Research and Quality. Initial risk factors were chosen based on clinical input and published research.

2b4.2 Analytic Method (*Describe methods and rationale for development and testing of risk model or risk stratification including selection of factors/variables)***:**
Initial development of the risk model was done by AHRQ for measure NQI #03. A regression model including these risk groupings was then developed. Selection of risk factors was based on clinical knowledge, those used in previous studies or research protocols 2-5 and the data that would be consistently available in administrative databases. These included: birthweight (in 250 gram intervals), gender, multiple gestation, and gender and birthweight interacted. Risk factors were selected using the logistic regression stepwise selection method.

2b4.3 Testing Results *(Statistical risk model: Provide quantitative assessment of relative contribution of model risk factors; risk model performance metrics including cross-validation discrimination and calibration statistics, calibration curve and risk decile plot, and assessment of adequacy in the context of norms for risk models. Risk stratification: Provide quantitative assessment of relationship of risk factors to the outcome and differences in outcomes among the strata)***:**
The c-statistic for the initially developed risk model was 0.744.

2b4.4 If outcome or resource use measure is not risk adjusted, provide rationale and analyses to justify lack of adjustment:
Not Applicable

2b5. Identification of Meaningful Differences in Performance. (*The performance measure scores were appropriately analyzed and discriminated meaningful differences in quality.*)

2b5.1 Data/Sample *(Describe the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)***:**
As previously noted the PC measure set has been in national use since the 2nd quarter of 2010. Demographics of organizations collecting and reporting data on these measures are as follows:
163 health care organizations representing various types, locations and sizes:
10 For Profit, 91 Not for Profit, 46 Military Facilities, 9 County, 2 State, 5 Other
15 >=500 beds; 29 250-499 beds; 50 100-249 beds; 69 <100 beds
Located in: AE, AK, AL, AP, AR, AZ, CA, DO, DC, FL, GA, HI, ID, IL, IN, KS, KY, LA, MA, MD, MI, MN, MO, MS, MT, NC, NE, NV, NY, OH,OK, PA, PR, RI, SC, TN, TX, VA, WA, WI, WV
26 performance measurement systems

2b5.2 Analytic Method *(Describe methods and rationale to identify statistically significant and practically/meaningfully differences in performance)***:**
The method used to analyze meaningful differences in performance at The Joint Commission is Target Analysis. The object of target analysis is to compare a health care organizations (HCO) data against a comparative norm for the purpose of evaluating performance improvement opportunities. When an organization's performance level is statistically significantly different from a comparative norm, it is considered a statistical deviation. A statistical deviation may be desirable or undesirable depending on the "direction of improvement" of the measure.
There are two components to the target analysis methodology used at The Joint Commission. Given the national average for a performance measure, a target range is constructed. Using generalized linear mixed models methodology (also known as hierarchical models), a predicted estimate of an HCO's performance, with a corresponding 95% confidence interval, is generated. This confidence interval is compared to the target range, to determine the HCOs' rating. The estimate of the organization's true performance is based on both the data from that organization and on data from the entire set of reporting organizations.

2b5.3 Results *(Provide measure performance results/scores, e.g., distribution by quartile, mean, median, SD, etc.; identification of statistically significant and meaningfully differences in performance)*:
PC-04 Distribution of Outliers

2011 1st Quarter Data:
Scores on this measure: N=79, Mean 1.78%, SD 0.1134
10th Percentile= 0%
25th Percentile= 0%
50th Percentile= 0%
75th Percentile= 0%
90th Percentile= 1.64%

79 (100%) Neutral – results not significantly different from target range

2b6. Comparability of Multiple Data Sources/Methods. (*If specified for more than one data source, the various approaches result in comparable scores.*)

2b6.1 **Data/Sample** *(Describe the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)*:
Multiple data sources are not used for this measure.

2b6.2 **Analytic Method** *(Describe methods and rationale for testing comparability of scores produced by the different data sources specified in the measure)*:
Not applicable

2b6.3 **Testing Results** *(Provide statistical results, e.g., correlation statistics, comparison of rankings; assessment of adequacy in the context of norms for the test conducted)*:
Not applicable

2c. Disparities in Care: H☐ M☐ L☐ I ☐ NA☐ (*If applicable, the measure specifications allow identification of disparities.*)

2c.1 **If measure is stratified for disparities, provide stratified results** *(Scores by stratified categories/cohorts)*: The measure is not stratified for disparities.

2c.2 **If disparities have been reported/identified (e.g., in 1b), but measure is not specified to detect disparities, please explain:**
Not Applicable

2.1-2.3 **Supplemental Testing Methodology Information:**

Steering Committee: Overall, was the criterion, *Scientific Acceptability of Measure Properties*, met?
(*Reliability and Validity must be rated moderate or high*) Yes☐ No☐
Provide rationale based on specific subcriteria:

If the Committee votes No, STOP

NEW (since 2012 endorsement): Data for Empirical Testing

2b2. VALIDITY TESTING

2b2.1. What level of validity testing was conducted? (*may be one or both levels*)

- ☐ **Critical data elements** (*data element validity must address ALL critical data elements*)
- ☐ **Performance measure score**
- ☒ **Empirical validity testing**
- ☐ **Systematic assessment of face validity of performance measure score as an indicator** of quality or resource use (*i.e., is an accurate reflection of performance on quality or resource use and can distinguish good from poor performance*)

2b2.2. For each level of testing checked above, describe the method of validity testing and what it tests (*describe the steps—do not just name a method; what was tested, e.g., accuracy of data elements compared to authoritative source, relationship to another measure as expected; what statistical analysis was used)*

ICD-9 to ICD-10 Conversion Process:
The goal was to convert ICD-9 to ICD-10 equivalent codes, consistent with the clinical intent of the original measure specifications. The Joint Commission worked with a certified coding expert throughout the conversion process. The 3M Coding Conversion Tool was utilized, including forward mapping of ICD-9 codes to ICD-10 codes as well as reverse mapping from ICD-10 to ICD-9 to ensure appropriateness. MSDRGs and instructions in the tabular index were also examined to ensure appropriate code mapping. Crosswalks comprising ICD-9

codes mapped to ICD-10 codes were created and reviewed by members of the Technical Advisory Panel, CMS subcontractors, and performance measurement system vendors prior to being posted for a 12 month public comment period. Feedback from the field indicated that the crosswalks generally were mapped correctly. Minor modifications to the code tables were made as needed. Final code tables were published in early 2015, well in advance of the mandated date of October 1, 2015.
Within the PC-Newborn population, there are two 2 subpopulations, i.e., Newborns with Blood Stream Infection or BSI, Newborns with Breast Feeding, each identified by Patient Age at admission and a specific group of diagnosis and procedure codes or lack thereof. The patients in each subpopulation are processed independently through each initial patient population flow. Patients may fall in both subpopulations depending on the presence or absence of the diagnosis codes or procedure codes and other data elements defined by the respective initial patient subpopulations **PC-04 -** Health Care-Associated Bloodstream Infections in Newborns which belongs to the population of the PC-Newborn BSI.

Patients admitted to the hospital for inpatient acute care are included in one of the PC Newborn subpopulations if they have:
Newborns with BSI - Patients with a Newborn Patient Age at admission (*Admission Date Birthdate*) ≤ 2 days **AND** satisfy conditions #1 through #3.

1. **NO** *ICD-9-CM Principal Diagnosis Code* as defined in Appendix A, Table 11.10.2,
2. **ONE** of the following:
    - an *ICD-9-CM Other Diagnosis Code* as defined in Appendix A, Tables 11.12, 11.13, 11.13.1, 11.14 Or *Birth Weight* >= 500g and <= 1499g
    - an *ICD-9-CM Other Diagnosis Code* as defined in Appendix A, Tables 11.15, 11.16, 11.16.1, 11.17 Or *Birth Weight* >=1500g with **ANY** OF THE FOLLOWING:
        - an *ICD-9-CM-Principal* or *Other Procedure Code* as defined in Appendix A, Tables 11.18 or 11.19
        - *Discharge Disposition* of 6 (expired) or a Missing *Discharge Disposition*
        - **NO** *ICD-9-CM Principal Diagnosis Code* as defined in Appendix A, Table 11.10.3
    - *Birth Weight* Missing or Unable To Determine (UTD).
3. **NO** *ICD-9-CM Other Diagnosis Code* as defined in Appendix A, Table 11.20 Or *Birth Weight* < 500g

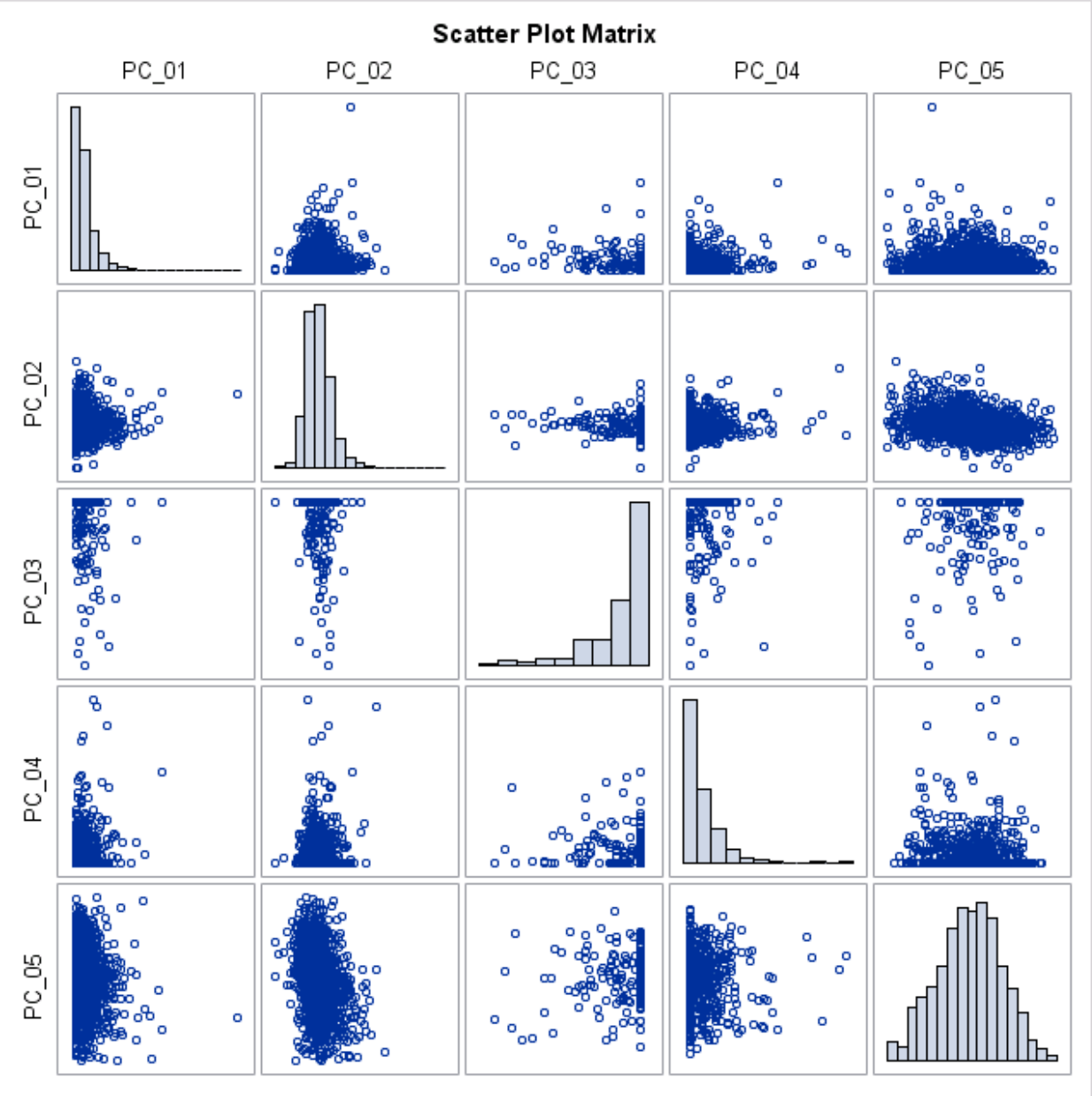
The data used to measure the validity of the PC measure are comprised of data from the third and fourth quarters of 2014, and the first and second quarters of 2015. 1,345 hospitals submitted 2,695,467 inpatient records for all the elected PC measures. The hospitals included in the analysis reported one year of data and had 30 or more denominator cases in the analysis period.
Measure convergent validity for PC-04 was assessed using hospitals patient level data from The Joint Commission warehouse. Measure specifications, including population identification, numerator and denominator statements, exclusions, and data elements and their definitions were found to be understandable, retrievable, and relevant in previous validity testing.

2b2.3. What were the statistical results from validity testing? (*e.g., correlation; t-test*)

Overall Descriptive statistics for sub population 2 - PC-04 Newborns with BSI
N=523 hospitals
n= 469,910 records submitted
Mean: 4.9%
Min = 0%
Percentile 10%: 0%
Percentile 25%: 0%
Median: 2.2%
Percentile 75%: 7.7%
Percentile 90%: 13.9%
Max = 66.1%

Simple Statistics						
Variable	N	Mean	Std Dev	Median	Minimum	Maximum
PC_01	1237	0.02753	0.03803	0.01734	0	0.51240
PC_02	1345	0.26287	0.07974	0.25410	0	1.00000
PC_03	162	0.97762	0.03311	0.99425	0.84615	1.00000
PC_04	523	0.05267	0.08432	0.02203	0	0.66129
PC_05	1352	0.49198	0.19284	0.50190	0.00317	1.00000

Spearman Correlation Coefficients Prob > |r| under H0: Rho=0 Number of Observations					
	PC_01	PC_02	PC_03	PC_04	PC_05
PC_01	1.00000 1237	0.06843 0.0163 1231	-0.26960 0.0006 159	0.10724 0.0169 496	-0.03538 0.2137 1237
PC_02	0.06843 0.0163 1231	1.00000 1345	-0.18318 0.0196 162	0.02807 0.5218 523	-0.32009 <.0001 1343
PC_03	-0.26960 0.0006 159	-0.18318 0.0196 162	1.00000 162	-0.03117 0.7030 152	0.07729 0.3283 162
PC_04	0.10724 0.0169 496	0.02807 0.5218 523	-0.03117 0.7030 152	1.00000 523	-0.03560 0.4165 523
PC_05	-0.03538 0.2137 1237	-0.32009 <.0001 1343	0.07729 0.3283 162	-0.03560 0.4165 523	1.00000 1352



The Spearman rank-order correlation is a nonparametric measure of association based on the ranks of the data values by measure PC-04 and hospitals. .We used this methodology because of the skewness of the distribution of the rates.

2b2.4. What is your interpretation of the results in terms of demonstrating validity? (i*.e., what do the results mean and what are the norms for the test conducted?*)

The correlation of PC-04 with PC-02 in the PC measure set is relatively weak but statistically significant. The correlation with the other PC measures is not significant. Although 90% of the hospital measure rates fall between 0 and 14%, there are still a number of hospitals with measure rates significantly greater than the median value of 2%, indicating that the performance of hospitals on this measure can be improved.

2b3. EXCLUSIONS ANALYSIS .
NA ☐ no exclusions — *skip to section* 2b4

2b3.1. Describe the method of testing exclusions and what it tests (*describe the steps—do not just name a method; what was tested, e.g., whether exclusions affect overall performance scores; what statistical*

analysis was used)

There were 469,910 admissions selected from the initial cohort. From among the 469,910 admissions in 523 hospitals, the descriptive statistics are given below.

The following exclusions were analyzed by subpopulation and measure for frequency and variability across providers:

- *ICD-9-CM Principal Diagnosis Code* for septicemias or bacteremia as defined in Appendix A, Table 11.10.2
- *ICD-9-CM Other Diagnosis Codes* for septicemias or bacteremia as defined in Appendix A, Table 11.10.2 or *ICD-9-CM Principal or Other Diagnosis Codes* for newborn septicemia or bacteremia as defined in Appendix A, Table 11.10 with a *Bloodstream Infection Present on Admission*
- *ICD-9-CM Other Diagnosis Codes* for birth weight < 500g as defined in Appendix A, Table 11.20 OR Birth Weight < 500g
- Enrolled in clinical trials
- Length of Stay < 2 days

2b3.2. What were the statistical results from testing exclusions? (*include overall number and percentage of individuals excluded, frequency distribution of exclusions across measured entities, and impact on performance measure scores*)

Exclusion Subpopulation 2 - PC-04

ICD-9-CM Principal Diagnosis Code for septicemias or bacteremia as defined in Appendix A **Exclusion:** Included in the initial population exclusion

ICD-9-CM Other Diagnosis Codes for septicemias or bacteremia as defined in Appendix A, Table 11.10.2 or *ICD-9-CM Principal or Other Diagnosis Codes* for newborn septicemia or bacteremia as defined in Appendix A, Table 11.10 with a *Bloodstream Infection Present on Admission*
Exclusion: Included in the initial population exclusion

Exclusion: Patients enrolled in clinical trials
Overall Occurrence n = 3,819
Overall Occurrence Percentage 0.8%
Minimum: 0%
10th Percentile: 0%
Median: 0%
90th Percentile: 0.7%
Maximum: 33%

Exclusion: Length of Stay < 2 days
Overall Occurrence n = 22,398
Overall Occurrence Percentage: 4.8%
Minimum: 0%
10th Percentile: 0%
Median: 4.2%
90th Percentile: 11.8%
Maximum: 82%

2b3.3. What is your interpretation of the results in terms of demonstrating that \exclusions are needed to prevent unfair distortion of performance results? (*i.e., the value outweighs the burden of increased data*

collection and analysis. *Note*: ***If patient preference is an exclusion****, the measure must be specified so that the effect on the performance score is transparent, e.g., scores with and without exclusion*)

The difference between the 10th and 90th percentiles of the distribution of exclusion rates is narrow indicating that the occurrence is random and likely would not bias performance results.

It is believed that all of the exclusions should be retained for the following reasons:

Exclusion: *ICD-9-CM Principal Diagnosis Code* for septicemias or bacteremias as defined in Appendix A, Table 11.10.2
Rationale: This is to remove all cases for newborns with principal diagnosis codes for infection on Table 11.10.2 who were already infected when they were received as transfers within the first 2 days of birth from another hospital.

Exclusion *ICD-9-CM Other Diagnosis Codes* for septicemias or bacteremias as defined in Appendix A, Table 11.10.2 or *ICD-9-CM Principal or Other Diagnosis Codes* for newborn septicemia or bacteremia as defined in Appendix A, Table 11.10 with a *Bloodstream Infection Present on Admission.*
Rationale: This is remove the same group of newborns above with other diagnosis codes for infection on Table 11.10.2 OR with principal diagnosis codes for infection on Table 11.10. Newborns born at the reporting hospital who were born with an infection are removed with other diagnosis codes for infection on Table 11.10. For both groups in addition to the infection codes, there must ALSO be documentation that the infection was present on admission which requires a medical record review. The reason for this is because coders will sometimes apply an infection code when there is documentation of a suspected infection even with negative blood cultures.

Exclusion: Patients who *Exclusion* Length of Stay < 2 days
Rationale: Included for this measure in order to harmonize with other CMS/Joint Commission aligned measures.

Exclusion: Patients enrolled in a Clinical Trial
Rationale: To capture patients not enrolled in clinical trials studying pregnant patients or newborns.

Risk Adjustment/SDS Section

1.8 What were the patient-level sociodemographic (SDS) variables that were available and analyzed in the data or sample used? For example, patient-reported data (e.g., income, education, language), proxy variables when SDS data are not collected from each patient (e.g. census tract), or patient community characteristics (e.g. percent vacant housing, crime rate).

The two SDS risk factors that were collected with the patient data are race and ethnicity. The race risk factor was used to create an “African American” and an “other non-white” risk factor for the purpose of this evaluation.

2b2. VALIDITY TESTING
2b2.1. What level of validity testing was conducted? (*may be one or both levels*)

☒ **Critical data elements** (*data element validity must address ALL critical data elements*)
☐ **Performance measure score**
☐ **Empirical validity testing**
☐ **Systematic assessment of face validity of performance measure score as an indicator** of quality or resource use (*i.e., is an accurate reflection of performance on quality or resource use and can distinguish good from poor performance*)

2b2.2. For each level of testing checked above, describe the method of validity testing and what it tests (*describe the steps—do not just name a method; what was tested, e.g., accuracy of data elements compared to authoritative source, relationship to another measure as expected; what statistical analysis was used)*

The correlation of the measure with other process and outcome measures in the PC measure set was done. The percentiles of the distribution of measure rates for all reporting hospitals were calculated to look at the spread of the measure rates.

2b2.3. What were the statistical results from validity testing? (*e.g., correlation; t-test*)

See the results in the validity testing section

2b2.4. What is your interpretation of the results in terms of demonstrating validity? (i*.e., what do the results mean and what are the norms for the test conducted?*)

See the results in the validity testing section

2b4. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES
If not an intermediate or health outcome, or PRO-PM, or resource use measure, skip to section 2b5.

2b4.1. What method of controlling for differences in case mix is used?

- ☐ **No risk adjustment or stratification**
- ☒ **Statistical risk model with 6 risk factors**
- ☐ **Stratification by** Click here to enter number of categories **risk categories**
- ☐ **Other,** Click here to enter description

2b4.2. If an outcome or resource use measure is not risk adjusted or stratified, provide rationale and analyses to demonstrate that controlling for differences in patient characteristics (case mix) is not needed to achieve fair comparisons across measured entities.

NA

2b4.3. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors (clinical factors or sociodemographic factors) used in the statistical risk model or for stratification by risk (*e.g., potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of p<0.10; correlation of x or higher; patient factors should be present at the start of care*)

Patient factors were initially selected based on clinical knowledge, those used in previous studies or research protocols, and the data that would be consistently available in administrative databases. One of the AHRQ Pediatric Quality indicators, NQI #3 – Neonatal Bloodstream Infection, was a similar measure already developed with a tested risk model. We initially used all the risk factors identified for the AHRQ risk model. Logistic regression was used to select patient factors using a combination of backward, forward and stepwise selection. A statistical significant of p<0.10 was used as a threshold in initially including a risk factor in the model. Risk factors significant at the p<0.01 level were included in the final model.

2b4.4a. What were the statistical results of the analyses used to select risk factors?
The following risk factors were all significant at p<0.0001:

Birth Weight 500 to 749g
Birth Weight 750 to 750g
Modified DRG Newborn Transfers Out or Died
Congenital Anomaly Gastrointestinal Anomaly
Congenital Anomaly Cardiovascular Anomaly
Out-born Birth Newborns Transfers In

2b4.4b. Describe the analyses and interpretation resulting in the decision to select SDS factors (e.g. prevalence of the factor across measured entities, empirical association with the outcome, contribution of unique variation in the outcome, assessment of between-unit effects and within-unit effects)

Race and ethnicity data were the SDS factors available in our database and were evaluated for significance by adding to the risk model. Risk factors for African American race other non-white race and Hispanic ethnicity were created and added to the risk model given above. African American race and Hispanic ethnicity were both found to be significantly significant in the model, after adjusting for the other risk factors, and increased the c-statistic of the risk model to 0.702.

2b4.5. Describe the method of testing/analysis used to develop and validate the adequacy of the statistical model or stratification approach (*describe the steps—do not just name a method; what statistical analysis was used*)

Risk model discrimination statistics and calibration curves were used to validate the adequacy of the risk model.

Provide the statistical results from testing the approach to controlling for differences in patient characteristics (case mix) below.
If stratified, skip to 2b4.9

2b4.6. Statistical Risk Model Discrimination Statistics (*e.g., c-statistic, R-squared*)**:**

The c-statistic for the final model was c=0.654

2b4.7. Statistical Risk Model Calibration Statistics (*e.g., Hosmer-Lemeshow statistic*):

The Hosmer-Lemeshow statistic for the final model was 28.4 which was significant (p<0.01)when compared to a chi-square distribution with 2 df.

2b4.8. Statistical Risk Model Calibration – Risk decile plots or calibration curves:

decile	Obs	Variable	Label	Mean	Std Dev	Minimum	Maximum
1		ind	observed rate	0.1516626	0.0282709	0.1117244	0.1837875
		pind	predicted rate	0.1574907	0.0556475	0.1029632	0.2487268
2		ind	observed rate	0.0835197	0.0133219	0.0659307	0.0973574
		pind	predicted rate	0.0819605	0.0129435	0.0657599	0.0940406
3		ind	observed rate	0.0617654	0.0013132	0.0608368	0.0626940
		pind	predicted rate	0.0567096	0.0040032	0.0538789	0.0595403
4		ind	observed rate	0.0519364	.	0.0519364	0.0519364
		pind	predicted rate	0.0489776	.	0.0489776	0.0489776
7		ind	observed rate	0.0326055	.	0.0326055	0.0326055
		pind	predicted rate	0.0337439	.	0.0337439	0.0337439

2b4.9. Results of Risk Stratification Analysis:
NA

2b4.10. What is your interpretation of the results in terms of demonstrating adequacy of controlling for differences in patient characteristics (case mix)? (i*.e., what do the results mean and what are the norms for the test conducted*)

The risk factors only had a moderate impact on risk due to the rarity of the risk factors in the patient population. Over 75% of the patients had none of the risk factors, making evaluation of the model difficult. The predicted probabilities ranged from 0.03 to 0.47, although only 5% of the predicted probabilities were over 0.10. For calibration the predicted rates were close to the observed, within 0.01.

3. Feasibility

Extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

3a. Byproduct of Care Processes

For clinical measures, the required data elements are routinely generated and used during care delivery (e.g., blood pressure, lab test, diagnosis, medication order).

3a.1. **Data Elements Generated as Byproduct of Care Processes.**

generated by and used by healthcare personnel during the provision of care, e.g., blood pressure, lab value, medical condition, Coded by someone other than person obtaining original information (e.g., DRG, ICD-9 codes on claims), Abstracted from a record by someone other than person obtaining original information (e.g., chart abstraction for quality measure or registry)

If other:

3b. Electronic Sources

The required data elements are available in electronic health records or other electronic sources. If the required data are not in electronic health records or existing electronic sources, a credible, near-term path to electronic collection is specified.

3b.1. **To what extent are the specified data elements available electronically in defined fields?** *(i.e., data elements that are needed to compute the performance measure score are in defined, computer-readable fields)*

Some data elements are in defined fields in electronic sources

3b.2. **If ALL the data elements needed to compute the performance measure score are not from electronic sources, specify a credible, near-term path to electronic capture, OR provide a rationale for using other than electronic sources.**

PC-04 is in the queue to be re-engineered as an eCQM as resources permit in the future.

3b.3. **If this is an eMeasure, provide a summary of the feasibility assessment in an attached file or make available at a measure-specific URL.**

Attachment:

3c. Data Collection Strategy

Demonstration that the data collection strategy (e.g., source, timing, frequency, sampling, patient confidentiality, costs associated with fees/licensing of proprietary measures) can be implemented (e.g., already in operational use, or testing demonstrates that it is ready to put into operational use). For eMeasures, a feasibility assessment addresses the data elements and measure logic and demonstrates the eMeasure can be implemented or feasibility concerns can be adequately addressed.

3c.1. **Describe what you have learned/modified as a result of testing and/or operational use of the measure regarding data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, time and cost of data collection, other feasibility/implementation issues.**

IF a PRO-PM, consider implications for both individuals providing PROM data (patients, service recipients, respondents) and those whose performance is being measured.

At the present time, hospitals using this performance measure generally collect measure data via manual review of the paper medical record, the EMR or a combination of both. Collected data are submitted to The Joint Commission on a quarterly basis, by way of contracted performance measurement system vendors, as described previously. Specifications for this measure are freely available to anyone who wishes to use the measure. Feedback from hospitals using this measure indicates that required data elements are generally available in the medical record, and measure specifications are robust and easy to understand. As described above, as feedback from measure users has indicated the need for clarification or revision of measure specifications, this has taken place. The Joint Commission added Vital Records as an additional data source in the current measure specifications.

3c.2. **Describe any fees, licensing, or other requirements to use any aspect of the measure as specified** *(e.g., value/code set, risk model, programming code, algorithm)*.

There are no fees or licensing requirements to use the Joint Commission performance measures, all of which are in the public domain.

4. Usability and Use

Extent to which potential audiences (e.g., consumers, purchasers, providers, policy makers) are using or could use performance results for both accountability and performance improvement to achieve the goal of high-quality, efficient healthcare for individuals or populations.

4a. Accountability and Transparency

Performance results are used in at least one accountability application within three years after initial endorsement and are publicly reported within six years after initial endorsement (or the data on performance results are available). If not in use at the time of initial endorsement, then a credible plan for implementation within the specified timeframes is provided.

4.1. Current and Planned Use

NQF-endorsed measures are expected to be used in at least one accountability application within 3 years and publicly reported within 6 years of initial endorsement in addition to performance improvement.

Planned	Current Use (for current use provide URL)
Public Reporting	Regulatory and Accreditation Programs Hospital Accreditation Program http://jointcommission.org Quality Improvement with Benchmarking (external benchmarking to multiple organizations) America's Hospitals: Improving Quality and Safety – The Joint Commission's Annual Report http://www.jointcommission.org/annualreport.aspx Quality Improvement (Internal to the specific organization) Perinatal Care Certification http://www.jointcommission.org/certification/perinatal_care_certification.aspx

4a.1. For each CURRENT use, checked above, provide:

- Name of program and sponsor
- Purpose
- Geographic area and number and percentage of accountable entities and patients included

- Name of program and sponsor Hospital Accreditation Program; The Joint Commission
- Purpose: An accreditation program that recognizes hospitals that meet standard requirements to provide safe and effective patient care.
- Geographic area and number and percentage of accountable entities and patients included Nationwide; 3300 Joint Commission-accredited hospitals (2014)

- Name of program and sponsor America's Hospitals: Improving Quality and Safety – The Joint Commission's Annual Report ; The Joint Commission
- Purpose: The annual report summarizes the performance of Joint Commission-accredited hospitals on 46 accountability measures of evidence-based care processes closely linked to positive patient outcomes, and provides benchmarks from Top Performer on Key Quality Measures® hospitals.
- Geographic area and number and percentage of accountable entities and patients included Nationwide; 3300 Joint Commission-accredited hospitals (2014)

- Name of program and sponsor Perinatal Care Certification; The Joint Commission
- Purpose: A certification program that recognizes hospitals that have achieved integrated, coordinated, patient-centered care for clinically uncomplicated pregnancies and births.
- Geographic area and number and percentage of accountable entities and patients included Nationwide; Ten Joint Commission-accredited hospitals (2015)

4a.2. If not currently publicly reported OR used in at least one other accountability application (e.g., payment program,

certification, licensing) what are the reasons? (*e.g., Do policies or actions of the developer/steward or accountable entities restrict access to performance results or impede implementation?*)
Not Applicable

4a.3. **If not currently publicly reported OR used in at least one other accountability application, provide a credible plan for implementation within the expected timeframes -- any accountability application within 3 years and publicly reported within 6 years of initial endorsement.** (*Credible plan includes the specific program, purpose, intended audience, and timeline for implementing the measure within the specified timeframes. A plan for accountability applications addresses mechanisms for data aggregation and reporting.*)
Not Applicable

4b. Improvement

Progress toward achieving the goal of high-quality, efficient healthcare for individuals or populations is demonstrated. If not in use for performance improvement at the time of initial endorsement, then a credible rationale describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.

4b.1. **Progress on Improvement. (Not required for initial endorsement unless available.)**
Performance results on this measure (current and over time) should be provided in 1b.2 and 1b.4. Discuss:

- **Progress (trends in performance results, number and percentage of people receiving high-quality healthcare)**
- **Geographic area and number and percentage of accountable entities and patients included**

Some progress have been made in eliminating health care-associated bloodstream infections in newborns based on 2014 Joint Commission ORYX performance measurement data which show hospitals in in the median, lower quartile and 10th percentile reporting no infections. A spike in the aggregate rate (3.2%) was noted with the addition of 1060 more hospitals reporting the data in 2014 (n=1218) when compared with a 2010 rate of 0.02% reported from 158 hospitals. Beginning with January 1, 2016 discharges, an additional 821 accredited hospitals will begin reporting the data. The new reporting requirement will capture approximately 80% of the accredited hospitals with maternity services in the US. As a result, the rates may potentially rise again

4b.2. **If no improvement was demonstrated, what are the reasons? If not in use for performance improvement at the time of initial endorsement, provide a credible rationale that describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.**
Not Applicable

4c. Unintended Consequences

The benefits of the performance measure in facilitating progress toward achieving high-quality, efficient healthcare for individuals or populations outweigh evidence of unintended negative consequences to individuals or populations (if such evidence exists).

4c.1. **Were any unintended negative consequences to individuals or populations identified during testing; OR has evidence of unintended negative consequences to individuals or populations been reported since implementation? If so, identify the negative unintended consequences and describe how benefits outweigh them or actions taken to mitigate them.**
Unintended Consequence:
Some hospitals reported an increase in the burden of data abstraction for the data element Birth Weight when cases were not routinely coded with diagnosis codes for birth weight.
Mitigating Action:
Vital Records reports, delivery logs and clinical information systems were added as acceptable data sources to help hospitals identify birth weights via reports to help reduce the burden of data abstraction.

Unintended Consequence:
Cases with a length of stay greater than 120 days with health care-associated bloodstream infections were being excluded due to the denominator exclusion.
Mitigating Action:
The length of stay > 120 days was removed from the denominator excluded population.

Unintended Consequence:
Some hospitals reported cases with infection codes were failing when bloodstream infections were present on admission.
Mitigating Action:
A new data element Bloodstream Infection Present on Admission was added to trigger a review to determine if the infection was present on admission in order to remove the case from the measure.

Unintended Consequence:
Some hospitals reported that cases coded with infections were failing when newborns experienced bloodstream infections that were not health care-associated later during the hospitalization, i.e., necrotizing enterocolitis, pneumonia, urosepsis, etc.
Mitigating Action:
A new data element Bloodstream Infection Confirmed was added as a final check at the numerator level for cases with infection codes in order to confirm if the infection was health care-associated in order to exclude cases due to other causes there were not health care-associated.

5. Comparison to Related or Competing Measures

If a measure meets the above criteria and there are endorsed or new related measures (either the same measure focus or the same target population) or competing measures (both the same measure focus and the same target population), the measures are compared to address harmonization and/or selection of the best measure.

5. Relation to Other NQF-endorsed Measures

Are there related measures (conceptually, either same measure focus or target population) or competing measures (conceptually both the same measure focus and same target population)? If yes, list the NQF # and title of all related and/or competing measures.

Yes

5.1a. **List of or competing measures (selected from NQF-endorsed measures)**

0304 : Late sepsis or meningitis in Very Low Birth Weight (VLBW) neonates (risk-adjusted)

0478 : Neonatal Blood Stream Infection Rate (NQI 03)

5.1b. **If related or competing measures are not NQF endorsed please indicate measure title and steward.**

5a. Harmonization

The measure specifications are harmonized with related measures;

OR

The differences in specifications are justified

5a.1. **If this measure conceptually addresses EITHER the same measure focus OR the same target population as NQF-endorsed measure(s):**

Are the measure specifications completely harmonized?

No

5a.2. **If the measure specifications are not completely harmonized, identify the differences, rationale, and impact on interpretability and data collection burden.**

Measure 0304 addresses infections in the newborn. Measure 0304 evaluates very low birth weight newborns for both late sepsis and meningitis with birth weights between 401 and 1500 Gms and a gestational age between 22 weeks 0 days and 28 weeks six days. Measure 0304 also evaluates all newborns who are in the hospital after 3 days of birth. Numerator inclusions for measure 0304 are a bacterial pathogen recovered from a blood culture and/or cerebrospinal fluid culture obtained after Day 3 of life OR all 3 of the following: 1.) Coagulase Negative Staphylococcus recovered from a blood culture from either a central line or peripheral blood sample and/or is recovered from cerebrospinal fluid by lumbar puncture, ventricular tap or ventricular drain 2.) One or more signs of generalized infection (i.e., apnea, temperature instability, feeding intolerance, worsening respiratory distress or hemodynamic instability) and 3.) Treatment with 5 or more days of intravenous antibiotics. The major differences between measure 0304 and measure 1731 are: • Measure 1731 does not include cases with meningitis based on results from cerebrospinal fluid cultures • Measure 1731 includes birth weights which are 500 Gms or more rather than 400 Gms or more, and measure 1731 also includes newborns 1500 gms or more with one or more specific medical indication: major surgery, mechanical ventilation, expired or transferred-in. • Measure 1731 excludes newborns born with infections within the first 48 hours of admission and newborns with bloodstream infections occurring after the first 48 hours after birth that are due to causes that are not health care-associated, i.e., necrotizing enterocolitis, urosepsis, etc.

5b. Competing Measures

The measure is superior to competing measures (e.g., is a more valid or efficient way to measure);

OR

Multiple measures are justified.

5b.1. **If this measure conceptually addresses both the same measure focus and the same target population as NQF-endorsed measure(s):**

Describe why this measure is superior to competing measures (e.g., a more valid or efficient way to measure quality); OR provide a rationale for the additive value of endorsing an additional measure. (Provide analyses when possible.)

Measure 0478 is similar to this measure. The fundamental differences are that measure 0478 has been developed to collect all data elements using administrative data. Such an approach has led in some cases to loss of specificity available through review of the medical record. The two measures have been harmonized to the extent possible; however, there are intrinsic differences which are addressed in a comparison table in the attachment found in Section A.1 Supplemental Materials.

Appendix

A.1 Supplemental materials may be provided in an appendix. All supplemental materials (such as data collection instrument or methodology reports) should be organized in one file with a table of contents or bookmarks. If material pertains to a specific submission form number, that should be indicated. Requested information should be provided in the submission form and required attachments. There is no guarantee that supplemental materials will be reviewed.

Attachment **Attachment:** NQI_3__TJC_comparison_2015_harmonization.docx

Contact Information

Co.1 **Measure Steward (Intellectual Property Owner):** The Joint Commission
Co.2 **Point of Contact:** Ann, Watt, awatt@jointcommission.org, 630-792-5944-
Co.3 **Measure Developer if different from Measure Steward:** The Joint Commission
Co.4 **Point of Contact:** Ann, Watt, awatt@jointcommission.org, 630-792-5944-

Additional Information

Ad.1 **Workgroup/Expert Panel involved in measure development**
Provide a list of sponsoring organizations and workgroup/panel members' names and organizations. Describe the members' role in measure development.
Michael Ross, MD, MPH (Chair)
Harbor-UCLA Medical Center
Torrance, CA

Wanda Barfield, MD, MPH
Centers for Disease Control and Prevention
Atlanta, GA

Kenneth E. Brown, MD, MBA, FACOG, FACHE
Woman´s Hospital
Lafayette, LA

Martin McCaffrey, MD
UNC North Carolina Children's Hospital
Chapel Hill, NC

Cathy Collins-Fulea, MSN, CNM
Henry Ford Hospital
Detroit, MI

Janet H. Muri, MBA
National Perinatal Information Center/
Quality Analytic Services
Providence, RI

Kathleen Simpson, PhD, RNC, FAAN
St. John's Mercy Medical Center
St. Louis, MO

Michael Socol, MD
Northwestern University Medical School
Chicago, IL

Rebecca Zimmermann, MPP
America's Health Insurance Plans
Washington, DC
The technical advisory panel (TAP) members determined priority areas that could be evaluated to improve care related to perinatal care during the development timeframe. After implementation, minor revisions, acknowledged by TAP representatives, were made

to improve clarity. Hospital feedback will be reviewed during the reliability testing phase of the project to assist the TAP in making the final measure recommendations.

Measure Developer/Steward Updates and Ongoing Maintenance

Ad.2 **Year the measure was first released:** 2010

Ad.3 **Month and Year of most recent revision:** 10, 2015

Ad.4 **What is your frequency for review/update of this measure?** Biannual

Ad.5 **When is the next scheduled review/update for this measure?** 02, 2016

Ad.7 **Disclaimers:**

Ad.8 **Additional Information/Comments:**

MEASURE WORKSHEET

This document summarizes the evaluation of the measure as it progresses through NQF's Consensus Development Process (CDP). The information submitted by measure developers/stewards is included after the Brief Measure Information, Preliminary Analysis, and Pre-meeting Public and Member Comments sections.

To navigate the links in the worksheet: Ctrl + click link to go to the link; ALT + LEFT ARROW to return

Brief Measure Information

NQF #: 2829
Measure Title: PC-01 Elective Delivery
Measure Steward: The Joint Commission
Brief Description of Measure: This measure assesses patients with elective vaginal deliveries or elective cesarean births at >= 37 and < 39 weeks of gestation completed. This measure is a part of a set of five nationally implemented measures that address perinatal care (PC-02: Cesarean Section, PC-03: Antenatal Steroids, PC-04: Health Care-Associated Bloodstream Infections in Newborns, PC-05: Exclusive Breast Milk Feeding). PC-01, Elective Delivery is one of two of the measures in this set that have been reengineered as eCQMs and are included in the EHR Incentive Program and Hospital Inpatient Quality Reporting Program.
Developer Rationale: A reduction in the number of non-medically indicated elective deliveries at >=37 to <39 weeks gestation will result in a substantial decrease in neonatal morbidity and mortality, as well as a significant savings in health care costs. In addition, the rate of cesarean deliveries should decrease with fewer elective inductions resulting in decreased length of stay and health care costs.

The measure will assist health care organizations (HCOs) to track non-medically indicated early term elective deliveries and reduce the occurrence.

Numerator Statement: Patients with elective deliveries by either:
- Medical induction of labor while not in labor prior to the procedure
- Cesarean birth while not in labor and with no history of a prior uterine surgery

Denominator Statement: The Denominator is patients who deliver newborns with >= 37 and < 39 weeks of gestation completed.
Denominator Exclusions: ICD-9-CM, ICD-10-CM, or SNOMED CT codes for conditions possibly justifying elective delivery prior to 39 weeks gestation.

Measure Type: Process
Data Source: Electronic Clinical Data, Electronic Clinical Data : Electronic Health Record, Electronic Clinical Data : Pharmacy
Level of Analysis: Facility, Population : National

IF Endorsement Maintenance – Original Endorsement Date: Most Recent Endorsement Date:

New Measure -- Preliminary Analysis

Criteria 1: Importance to Measure and Report

1a. Evidence

1a. Evidence. The evidence requirements for a *process or intermediate outcome* measure is that it is based on a systematic review (SR) and grading of the body of empirical evidence where the specific focus of the evidence matches what is being measured.

This measure is the new eMeasure version of measure 0469. The information provided for Evidence and

Opportunity for Improvement is identical to that submitted for 0469. Measure 0469 will be discussed first – the ratings for evidence and opportunity for improvement will automatically be assigned to this eMeasure without further discussion.

The developer provides the following evidence for this measure:

- **Systematic Review of the evidence specific to this measure?** ☒ **Yes** ☐ **No**
- **Quality, Quantity and Consistency of evidence provided?** ☒ **Yes** ☐ **No**
- **Evidence graded?** ☒ **Yes** ☐ **No**

Summary of prior review in 2012

- The developer reported that "both ACOG and AAP have had guidelines in place for a number of years which do not support non-medically indicated elective deliveries at< 39 weeks gestation. Several studies consistently document increased morbidity associated with elective delivery before 39 weeks. The studies note that elective deliveries performed at < 39 weeks carry significant risk for the newborn (odds ratios 2.0-3-0 compared to newborns born between 39-41 weeks). " The previous Perinatal Committee agreed that the "evidence is strong that elective delivery prior to 39 weeks impacts newborn adversely."

Changes to evidence from last review

☒ **The developer attests that there have been no changes in the evidence since the measure was last evaluated.**

☐ **The developer provided updated evidence for this measure:**

Updates:

- In 2013 ACOG reaffirmed the Practice Bulletin for Induction of Labor (August 2009).
- An ACOG Committee Opinion from April 2013 (reaffirmed in 2015) reviewed the evidence and provided guidance for "Nonmedically Indicated Early-Term Deliveries" concluding that "Although there are specific indications for delivery before 39 weeks of gestation, a nonedically indicated early-term delivery is not appropriate."

Exception to evidence NA

Questions for the Committee:

The developer attests the underlying evidence for the measure #0469 (chart-abstracted version) has not changed since the last NQF endorsement review. Does the Committee agree the evidence basis for the chart-abstracted measure has not changed and can be applied to the eMeasure, therefore there is no need for repeat discussion and voting on Evidence?

Preliminary rating for evidence: ☒ **Pass** ☐ **No Pass**

1b. Gap in Care/Opportunity for Improvement and 1b. Disparities
Maintenance measures – increased emphasis on gap and variation

1b. Performance Gap. The performance gap requirements include demonstrating quality problems and opportunity for improvement.

No data specific to the eMeasure is submitted.

The developers provide data for the past several years:
Baseline 2Q 2010: 164 hospitals, 11,843 cases; national aggregate rate = 18.8%; 50th %tile = 15.5%
Beginning 2014, all hospitals with >110 births/year were required to report.

	2011	2012	2013	2014
# hospitals	166	170	200	1388
# patients	13,907	13,404	14,880	130,882
National aggregate rate	13.6%	8.0%	4.4%	3.3%
10^{th}-25^{th} -90^{th} %tile	1.5 -5 -9.8– 31.5%	0– 2.6-4.9-21.2%	0- 0- 2.6-13.9%	0-0 - 2.1 -8.7 %
Mean hospital rate (SD)	0.13998 (0.13183)	0.08296 (0.09555)	0.05737 (0.10193)	0.82(0.31)

Beginning in January 2016 hospitals with >300 births/year are required to report – an additional 821 (approximately 80% of all birthing hospitals)

Disparities
Data from use of this measure to understand any disparities is not provided. The developers refer to the literature that has reported increased rates of elective inductions for all races though higher rates are seen in non-Hispanic whites. Bailey, et al. (2014) recently reviewed data for 638 rural women, recruited prenatally from three counties in rural southern Appalachia, who delivered electively at = 37 weeks. Those delivered electively at early term were 7.7 times more likely to be low birth weight, 4.4 times more likely to have a neonatal intensive care unit admission, and 2.5 times more likely to develop jaundice. Patients living furthest from the hospital were most likely to deliver electively at <39 weeks.

Questions for the Committee:
- *Is there a gap in care that warrants a national performance eMeasure?*
- *Is this measure useful in understanding disparities in provision of healthcare to pregnant women?*

Preliminary rating for opportunity for improvement: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 1: Importance to Measure and Report (including 1a, 1b, 1c)

1a. Evidence to Support Measure Focus
Comments:
Process measure—eMeasure. Directly related to outcomes.

1b. Performance Gap
Comments:
Yes there is performance data provided. Improvement is noted however nationally there is still continued improvement that needs to occur. Data on the population and subgroups was not provided

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability

2a1. Reliability Specifications
Maintenance measures – no change in emphasis – specifications should be evaluated the same as with new measures

2a1. Specifications requires the measure, as specified, to produce consistent (reliable) and credible (valid) results about

the quality of care when implemented.

Data source(s): EHR

Specifications: HQMF specifications are provided – see technical review

- The developers report that measures "included in the EHR Incentive Program for Eligible Hospitals undergo an annual update to revise specifications based on updated research and clinical information or standards changes. Changes have been made to the eCQM specifications in order to reflect revisions to the chart abstracted measure from which this measure is derived."

eMeasure Technical Advisor review:

Submitted measure is an HQMF compliant eMeasure	The submitted eMeasure specifications follow the industry accepted format for eMeasure (HL7 Health Quality Measures Format (HQMF)). **HQMF specifications** ☒ **Yes** ☐ **No**
Documentation of HQMF or QDM limitations	N/A – All components in the measure logic of the submitted eMeasure are represented using the HQMF and QDM
Value Sets	The submitted eMeasure specifications uses existing value sets when possible and uses new value sets that have been vetted through the VSAC
Measure logic is unambiguous	Submission includes test results from a simulated data set demonstrating the measure logic can be interpreted precisely and unambiguously Demonstrated through Bonnie testing
Feasibility Testing	The submission contains a feasibility assessment that addresses data element feasibility and follow-up with the measure developer indicates that the measure logic is feasible. This is a legacy eMeasure included in the Meaningful Use program. The developer submitted Bonnie testing results to establish feasibility, and provided an interpretation of the testing process and results in the measure testing attachment

Questions for the Committee :

- *Are the eMeasure specification fully aligned with the chart-based measure 0469?*

2a2. Reliability Testing Testing attachment
Maintenance measures – less emphasis if no new testing data provided

2a2. Reliability testing demonstrates if the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise enough to distinguish differences in performance across providers.

SUMMARY OF TESTING – see data element validity testing

Reliability testing level ☐ **Measure score** ☒ **Data element** ☐ **Both**

Reliability testing performed with the data source and level of analysis indicated for this measure ☒ **Yes** ☐ **No**

Method(s) of reliability testing - Data element validity testing will also count for reliability -see validity section

Results of reliability testing - see validity section

Preliminary rating for reliability: ☐ **High** ☐ **Moderate** ☐ **Low** ☐ **Insufficient**
The rating for reliability will be the same as validity.

2b. Validity

2b1. Validity: Specifications

2b1. Validity Specifications. This section should determine if the measure specifications are consistent with the evidence.

Specifications consistent with evidence in 1a. ☒ **Yes** ☐ **Somewhat** ☐ **No**

Question for the Committee:

- *Are the specifications consistent with the evidence?*

2b2. Validity testing

2b2. Validity Testing should demonstrate the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality.

SUMMARY OF TESTING

Validity testing level ☐ **Measure score** ☒ **Data element testing against a gold standard** ☐ **Both**

Method of validity testing of the measure score:

- ☐ **Face validity only**
- ☐ **Empirical validity testing of the measure score**

NQF supports the development and use of eMeasures. Until more widespread use of eMeasures provides data to formally test eMeasures, NQF will accept testing in a simulated dataset, such as the BONNIE tool, to meet the requirements for measure testing.

Validity testing method:

- BONNIE testing -simulated data set of 51 patients – see attachment
- The developer further describes the BONNIE testing "Cases were used to test the validity of each data element and timing relationship in the measure. Patient characteristics such as age, diagnosis, and length of stay were pre-determined to provide a variety of scenarios that adequately tested for patients passing each data element and failing each data element. Data included in cases and tested for this measure included diagnoses, differing types of delivery and uterine procedures, gestational ages, medications, patient orders, age, length of stay, and clinical observations, such as time of delivery."

Validity testing results:

- The developer reports that "All 51 cases passed or failed as expected based on the data included in the case, confirming the measure logic is accurate and valid."
- "Bonnie testing includes negative and positive testing of each data element in the measure. Positive testing ensures patients expected to be included in the measure are included. Negative testing ensures that patients who do not meet the data criteria are not included in the measure."

Questions for the Committee:

- *Is the test sample adequate to generalize for widespread implementation?*
- *Do the results demonstrate sufficient validity so that conclusions about quality can be made?*
- *Do you agree that the score from this measure as specified is an indicator of quality?*
- *Other specific question of the validity testing?*

2b3-2b7. Threats to Validity

2b3. Exclusions:

- The developer states the "Exclusions in the eMeasure align with the chart-based version of the measure, and are clinically necessary for the interpretation of the measure. No additional exclusion analyses were done on the eMeasure."

Questions for the Committee:

- *Are the exclusions consistent with the evidence?*
- *Is it reasonable to assume that the impact of exclusions will be similar for the eMeasure and the chart-abstracted measure?*

2b4. Risk adjustment: **Risk-adjustment method** ☒ **None** ☐ **Statistical model** ☐ **Stratification**

2b5. Meaningful difference *(can statistically significant and clinically/practically meaningful differences in performance measure scores can be identified)*:

- The developer submitted additional information on meaningful differences and also referred to results from 0469.

2b6. Comparability of data sources/methods: No comparison testing is available.

2b7. Missing Data

- The developer reports "Data not present in the structured field from which the measure draws will not be included in the measure calculation. In the Bonnie testing environment, missing data are tested as an expected "Fail" for that data element, and actual performance (whether or not the case fails) is compared to expectations to ensure missing data impacts data element and overall measure calculation as expected.

Guidance from algorithm: Specifications consistent with evidence (Box 1) → threats to validity empirically assessed (Box 2) →empirical testing of data elements using BONNIE tool →acceptable → Moderate

Preliminary rating for validity: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 2: Scientific Acceptability of Measure Properties (including all 2a, 2b, and 2d)

2a1. & 2b1. Specifications

Comments:

Clearly defined. Codes are included, but could always add more. Won't be able to include every potential justified reason for an early delivery. No concerns about consistent implementation.

Specifications are consistent with the evidence.

2a2. Reliability Testing

Comments:

Efficient sample size and appropriate for widespread implementation.

2b2. Validity Testing

Comments:

The sample size used for validity testing with the Bonnie test was insufficient (51).

2b3. Exclusions Analysis

2b4. Risk Adjustment/Stratification for Outcome or Resource Use Measures

2b5. Identification of Statistically Significant & Meaningful Differences In Performance

2b6. Comparability of Performance Scores When More Than One Set of Specifications

2b7. Missing Data Analysis and Minimizing Bias

Comments:

Missing data is a rarity in this population and should not affect the overall score.

Criterion 3. Feasibility

Maintenance measures – no change in emphasis – implementation issues may be more prominent

3. Feasibility is the extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

- This is an eMeasure for use in EHRs.
- The developer summarizes "Cases were used to test the validity of each data element and timing relationship in the measure. Patient characteristics such as age, diagnosis, and length of stay were pre-determined to provide a variety of scenarios that adequately tested for patients passing each data element and failing each data element. Data included in cases and tested for this measure included diagnoses, differing types of delivery and uterine procedures, gestational ages, medications, patient orders, age, length of stay, and clinical observations, such as time of delivery."
- BONNIE testing supports the feasibility of this measure.

Questions for the Committee:

- *Are the required data elements routinely generated and used during care delivery?*
- *Are the required data elements available in electronic form, e.g., EHR or other electronic sources?*
- *Is the data collection strategy ready to be put into operational use?*
- *Does the eMeasure Feasibility analysis demonstrate acceptable feasibility in multiple EHR systems and sites?*

Preliminary rating for feasibility: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 3: Feasibility

3a. Byproduct of Care Processes

3b. Electronic Sources

3c. Data Collection Strategy

Comments:

Data elements are routinely generated and should be available in electronic form for facilities with an EMR. Facilities without an EMR could not report this measure.

Criterion 4: Usability and Use

Maintenance measures – increased emphasis – much greater focus on measure use and usefulness, including both impact /improvement and unintended consequences

4. Usability and Use evaluate the extent to which audiences (e.g., consumers, purchasers, providers, policymakers) use or could use performance results for both accountability and performance improvement activities.

Current uses of the measure

Publicly reported? ☒ **Yes** ☐ **No**

Current use in an accountability program? ☒ **Yes** ☐ **No**
OR
Planned use in an accountability program? ☐ **Yes** ☐ **No**

Accountability program details

- Payment Program --Hospital Inpatient Quality Reporting Program https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/HospitalQualityInits/HospitalRHQDAPU.html

- Regulatory and Accreditation Programs Hospital Accreditation Program http://jointcommission.org
- EHR Incentive Program https://www.cms.gov/Regulations-and-Guidance/Legislation/EHRIncentivePrograms/index.html?redirect=/ehrincentiveprograms

Improvement results Data specific to the eMeasure is not available. See measure 0469.

Unexpected findings (positive or negative) during implementation
No information was provided.

Potential harms
No harms were identified.

Feedback:

***Questions for the Committee**:*
- *How can the performance results be used to further the goal of high-quality, efficient healthcare?*
- *Do the benefits of the measure outweigh any potential unintended consequences?*

Preliminary rating for usability and use: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments
Criteria 4: Usability and Use

4a. Accountability and Transparency
4b. Improvement
4c. Unintended Consequences
Comments:
TJC, CMS, Value Based Purchasing.

Criterion 5: Related and Competing Measures

Related or competing measures
This is the eMeasure version of #0469.

Harmonization

Pre-meeting public and member comments

-

1. IMPACT, OPPORTUITY, EVIDENCE - IMPORTANCE TO MEASURE AND REPORT

Importance to Measure and Report is a threshold criterion that must be met in order to recommend a measure for endorsement. All three subcriteria must be met to pass this criterion. See guidance on evidence.

Measures must be judged to be important to measure and report in order to be evaluated against the remaining criteria. **(evaluation criteria)**

1c.1 Structure-Process-Outcome Relationship *(Briefly state the measure focus, e.g., health outcome, intermediate clinical outcome, process, structure; then identify the appropriate links, e.g., structure-process-health outcome; process-health outcome; intermediate clinical outcome-health outcome)***:**

The focus of the measure is to decrease the number of elective deliveries >> population determined >> population assessed >> patient delivers spontaneously or planned delivery greater or equal to 39 weeks gestation >> improved maternal and fetal outcomes >> decreased length of stay and fetal morbidity and mortality.

1c.2-3 Type of Evidence *(Check all that apply)***:**

Clinical Practice Guideline

Systematic review of body of evidence (other than within guideline development)

1c.4 Directness of Evidence to the Specified Measure *(State the central topic, population, and outcomes addressed in the body of evidence and identify any differences from the measure focus and measure target population)***:**

The central topic for the measure is the reduction of elective deliveries at >= 37 and < 39 weeks of gestation completed. The target population for the performance measure is consistent with the body of evidence supporting the reduction of elective deliveries.

1c.5 Quantity of Studies in the Body of Evidence *(Total number of studies, not articles*)**:** No randomized-control trials (RCTs) were identified for early-term elective deliveries. RCTs were only identified for post-term elective deliveries versus expectant management. Given the current amount of population data available on the harms of early term and late pre-term delivery, it would be unethical to conduct such a study. Several studies were identified which were retrospective cohort or prospective observational in design examining thousands of births and the potential for adverse outcomes for both mother and newborn. In addition, several recent studies were identified addressing quality improvement interventions that were successful in reducing non-medically indicated early term elective deliveries.

1c.6 Quality of Body of Evidence (*Summarize the certainty or confidence in the estimates of benefits and harms to patients across studies in the body of evidence resulting from study factors. Please address: a) study design/flaws; b) directness/indirectness of the evidence to this measure (e.g., interventions, comparisons, outcomes assessed, population included in the evidence); and c) imprecision/wide confidence intervals due to few patients or events):* The quality of evidence supporting the reduction in the number of non-medically indicated elective deliveries is moderate . It is noteworthy to examine the fact that randomized control trials cannot be conducted, as one cannot randomly select women to agree to an elective delivery at < 39 weeks gestation.

As previously noted, both ACOG and AAP have had guidelines in place for a number of years which do not support non-medically indicated elective deliveries at > 39 weeks gestation. Several studies consistently document increased morbidity associated with elective delivery before 39 weeks. The studies note that elective deliveries performed at < 39 weeks carry

significant risk for the newborn (odds ratios 2.0-3-0 compared to newborns born between 39-41 weeks).

In spite of the fact that all studies reviewed were either retrospective or prospective cohort studies, no study design flaws were noted.

1c.7 **Consistency of Results <u>across Studies</u>** *(Summarize the consistency of the magnitude and direction of the effect):* The body of evidence consistently supports the benefit of reduction of non-medically indicated early term elective deliveries. All studies show an increase in the number of neonatal morbidities associated with early term deliveries, subsequent reduction of elective non-medically indicated deliveries reduces harm to the neonate. All studies demonstrated similar findings related to the direction of effect, though the magnitude varied from study to study, i.e., 8-17.8% increase in NICU admissions, rates of adverse respiratory outcomes, mechanical ventilation, newborn sepsis, hypoglycemia, admission to the NICU and hospitalization of 5 days or more increased by a factor of 1.8 to 4.2. and the incidence of transient tachypnea of the newborn, respiratory distress syndrome (RDS) and persistent pulmonary hypertension of the newborn were 3.1%, 0.25% and .17% respectively.

1c.8 **Net Benefit** *(Provide estimates of effect for benefit/outcome; identify harms addressed and estimates of effect; and net benefit - benefit over harms)***:**

As described before, elective deliveries performed at =>39 weeks gestation results in improved maternal and neonatal outcomes and will result in a substantial decrease in cesarean sections and neonatal morbidity, as well as substantial savings in health care costs. A recent study showed that by waiting until 39 weeks gestation, the NICU admissions fell from 12.8% to 5.9%, RDS fell from 3.7% to 0.9%, newborn sepsis fell from 7.0% to 2.5% and hospitalization > 5 days fell from 9.1% to 3.6%. This same study estimated that one-half million newborn intensive care unit days could be avoided in the U.S. population were a national rate of 1.7% to be achieved, with cost savings approaching $1 billion annually.

1c.9 **Grading of Strength/Quality of the Body of Evidence**. Has the body of evidence been graded? No

1c.10 **If body of evidence graded, identify the entity that graded the evidence including balance of representation and any disclosures regarding bias:** Not Applicable

1c.11 **System Used for Grading the Body of Evidence:** Other

1c.12 **If other, identify and describe the grading scale with definitions:** Although grading of the evidence was not determined during our systematic review, it was determined that the guideline developers accounted for a balanced representation of information, looked beyond one specialty group or discipline, and provided information that was accessible and met the requirements set out in this measure maintenance form.

1c.13 **Grade Assigned to the Body of Evidence:** Not Applicable

1c.14 **Summary of Controversy/Contradictory Evidence:** There is no documented evidence regarding controversy related to the reduction of non-medically indicated early term elective deliveries. A review of recent studies also supports the use of quality improvement interventions to further reduce the number of such deliveries.

1c.15 Citations for Evidence other than Guidelines*(Guidelines addressed below)*:

- American Academy of Family Physicians. (2000). Tips from Other Journals: Elective induction doubles cesarean delivery rate, 61, 4.Retrieved September 16, 2011 at: http://www.aafp.org/afp/20000215/tips/39.html.
- American College of Obstetricians and Gynecologists. (November 1996). ACOG Educational Bulletin.
- American College of Obstetricians and Gynecologists (ACOG) Practice Bulletin No 107. (2009). Induction of labor. Obstetrics & Gynecology. 114(2). 386-97.
- Clark, S., Miller, D., Belfort, M., Dildy, G., Frye, D., & Meyers, J. (2009). Neonatal and maternal outcomes associated with elective delivery. [Electronic Version]. Am J Obstet Gynecol. 200:156.e1-156.e4.
- Clark, S., Frye, D., Meyers, J., Belfort, M., Dildy, G., Kofford, S et al. (2010). Reduction in elective delivery at <39 weeks of gestation: comparative effectiveness of 3 approaches to change and the impact on neonatal intensive care admission and stillbirth. Am J Obstet Gynecol. 203:449.e1-6.
- Davidoff, M., Dias, T., Damus, K., Russell, R., Bettegowda, V.R., Dolan, S., et al. (2006). Changes in the gestational age distributin among U.S. singleton births: impacts on rates of late preterm birth, 1992-2002. Semin Perinatol. Feb;30(1):8-15.
- Engle, W.A. & Kominiarek, M.A. (2008). Late preterm infants, early term infants, and timing of elective deliveries. Clin Perinatol. 35:325-41.
- Glantz, J. (Apr.2005). Elective induction vs. spontaneous labor associations and outcomes. [Electronic Version]. J Reprod Med. 50(4):235-40.
- Tita, A., Landon, M., Spong, C., Lai, Y., Leveno, K., Varner, M, et al. (2009). Timing of elective repeat cesarean delivery at term and neonatal outcomes. [Electronic Version]. NEJM. 360:2, 111-120.

1c.16 Quote verbatim, <u>the specific guideline recommendation</u> *(Including guideline # and/or page #)*:

Clinical Considerations and Recommendations Page 389

What are the indications and contraindications to induction of labor?

Indications for induction of labor are not absolute but should take into account maternal and fetal conditions, gestational age, cervical status, and other factors. Following are examples of maternal or fetal conditions that may be indications for induction of labor:

- Abruptio placentae
- Chorioamnionitis
- Fetal demise
- Gestational hypertension
- Preeclampsia, eclampsia
- Premature rupture of membranes
- Postterm pregnancy
- Maternal medical conditions (eg, diabetes mellitus, renal disease, chronic pulmonary disease, chronic hypertension, antiphospholipid syndrome)

• Fetal compromise (eg, severe fetal growth restriction,

isoimmunization, oligohydramnios)

Labor also may be induced for logistic reasons, for example, risk of rapid labor, distance from hospital, or psychosocial indications. In such circumstances, at least one of the gestational age criteria in the box should be met, or fetal lung maturity should be established. A mature fetal lung test result before 39 weeks of gestation, in the absence of appropriate clinical circumstances, is not an indication for delivery. The individual patient and clinical situation should be considered in determining when induction of labor is contraindicated. Generally, the contraindications to labor induction are the same as those for spontaneous labor and vaginal delivery. They include, but are not limited to, the following situations:

• Vasa previa or complete placenta previa

• Transverse fetal lie

• Umbilical cord prolapse

• Previous classical cesarean delivery

• Active genital herpes infection

• Previous myomectomy entering the endometrial cavity

What criteria should be met before the cervix is ripened or labor is induced?

Assessment of gestational age and consideration of any potential risks to the mother or fetus are of paramount importance for appropriate evaluation and counseling before initiating cervical ripening or labor induction. The patient should be counseled regarding the indications for

induction, the agents and methods of labor stimulation, and the possible need for repeat induction or cesarean delivery. Although prospective studies are limited in evaluating the benefits of elective induction of labor, nulliparous women undergoing induction of labor with

unfavorable cervices should be counseled about a twofold increased risk of cesarean delivery (Level II-2). In addition, labor progression differs significantly for women with an elective induction of labor compared with women who have spontaneous onset of labor (Level II-2).

Allowing at least 12–18 hours of latent labor before diagnosing a failed induction may reduce the risk of cesarean delivery (Level II-2, 3). Additional requirements for cervical ripening and

induction of labor include assessment of the cervix, pelvis, fetal size, and presentation. Monitoring FHR and uterine contractions is recommended as for any high-risk patient in active labor. Although trained nursing personnel can monitor labor induction, a physician capable

of performing a cesarean delivery should be readily available.

1c.17 Clinical Practice Guideline Citation: • American College of Obstetricians and Gynecologists (ACOG) Practice Bulletin No 107. (2009). Induction of labor. Obstetrics & Gynecology. 114(2). 386-97.

1c.18 National Guideline Clearinghouse or other URL:

1c.19 Grading of Strength of Guideline Recommendation. Has the recommendation been graded? Yes

1c.20 If guideline recommendation graded, identify the entity that graded the evidence including balance of representation and any disclosures regarding bias: The American College of Obstericians and Gynecologists

1c.21 System Used for Grading the Strength of Guideline Recommendation: USPSTF

1c.22 If other, identify and describe the grading scale with definitions:

1c.23 Grade Assigned to the Recommendation: Level II

1c.24 Rationale for Using this Guideline Over Others: The American College of Obstetricians and Gynecologists the nation´s leading group of professionals providing health care for women. Practice Bulletins provide obstetricians and gynecologists with current information on established techniques and clinical management guidelines. The American College of Obstetricians and Gynecologists (the College) continuously surveys the field for advances to be incorporated in these series and monitors existing bulletins to ensure they are current. Individual bulletins are withdrawn from and added to the series on a continuing basis and reaffirmed periodically.

Based on the NQF descriptions for rating the evidence, what was the <u>developer's assessment</u> of the quantity, quality, and consistency of the body of evidence?

1c.25 Quantity: High **1c.26** Quality: Moderate**1c.27** Consistency**:** High

1. Evidence, Performance Gap, Priority – Importance to Measure and Report

Extent to which the specific measure focus is evidence-based, important to making significant gains in healthcare quality, and improving health outcomes for a specific high-priority (high-impact) aspect of healthcare where there is variation in or overall less-than-optimal performance. ***Measures must be judged to meet all subcriteria to pass this criterion and be evaluated against the remaining criteria.***

1a. Evidence to Support the Measure Focus – See attached Evidence Submission Form
0469_Evidence_MSF5.0_Data-635827510935417126.doc

1b. Performance Gap

Demonstration of quality problems and opportunity for improvement, i.e., data demonstrating:

- considerable variation, or overall less-than-optimal performance, in the quality of care across providers; and/or
- disparities in care across population groups.

1b.1. **Briefly explain the rationale for this measure** *(e.g., the benefits or improvements in quality envisioned by use of this measure)*
A reduction in the number of non-medically indicated elective deliveries at >=37 to <39 weeks gestation will result in a substantial decrease in neonatal morbidity and mortality, as well as a significant savings in health care costs. In addition, the rate of cesarean deliveries should decrease with fewer elective inductions resulting in decreased length of stay and health care costs.

The measure will assist health care organizations (HCOs) to track non-medically indicated early term elective deliveries and reduce the occurrence.

1b.2. **Provide performance scores on the measure as specified (current and over time) at the specified level of analysis.** *(This is required for endorsement maintenance. Include mean, std dev, min, max, interquartile range, scores by decile. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included). This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*
This measure is a legacy eCQM that is currently included in the Hospital Inpatient Quality Reporting Program (HIQR) and the EHR Incentive Program. At present, no performance data are yet available.

In CY2016, CMS is requiring organizations participating in HIQR to electronically submit 1 quarter of data on 4 of 28 available eCQMs. This measure is one of the 28.

For more information, refer to CY2016 IPPS Final Rule, located here: https://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/AcuteInpatientPPS/FY2016-IPPS-Final-Rule-Home-Page.html

1b.3. **If no or limited performance data on the measure as specified is reported in 1b2, then provide a summary of data from the literature that indicates opportunity for improvement or overall less than optimal performance on the specific focus of measurement.**
Early term elective deliveries are still being performed; however, the performance gap has continued to narrow over time. A goal of 5% or less based on recommendations from the PC Technical Advisory Panel (TAP) should be achievable. The chart-abstracted Perinatal Care (PC) core measures were added as a new core measure set in 2010 for hospitals to select in order to meet their ORYX performance measurement requirement for Joint Commission accreditation purposes. Approximately 164 hospitals reported the data with an average measure rate of 18.8% (n=11,843 patients). In January

2014, The Joint Commission required mandatory reporting of the chart-abstracted PC measure set for all accredited hospitals with 1100 births or more annually. 1388 hospitals reported the data with an average rate of 3.4% (n=130,882 patients). It is important to note that a performance gap of 3.7% exists for the 90th percentile of hospitals performing at 8.7%. The threshold for mandatory reporting was recently lowered to 300 births annually effective January 2016. The new reporting requirement will now capture approximately 80% of all accredited birthing hospitals. As a result, the rates may increase with the addition of approximately 821 more hospitals reporting data. Below is the specified level of analysis for PC-01 beginning with discharges April 1, 2010 through December 31, 2014.
2Q 2010: 11,843 denominator cases; 2,231 numerator cases; 164 hospitals; 18.8% national aggregate rate; 0.17827 mean of hospital rates; 0.12745 standard deviation; 33.3% 90th percentile rate; 23.7% 75th percentile rate/upper quartile; 15.5% 50th percentile rate/median rate; 9.0% 25th percentile rate/lower quartile; and 4.7% 10th percentile rate.
CY 2011: 1,3907 denominator cases; 1,892 numerator cases; 166 hospitals; 13.6% national aggregate rate; 0.13998 mean of hospital rates; 0.13183 standard deviation; 31.5% 90th percentile rate; 18.3% 75th percentile rate/upper quartile; 9.8% 50th percentile rate/median rate; 5% 25th percentile rate/lower quartile; and 1.5% 10th percentile rate.
CY 2012: 1,3404 denominator cases; 1,081 numerator cases; 170 hospitals; 8.0% national aggregate rate; 0.08296 mean of hospital rates; 0.09555 standard deviation; 21.2% 90th percentile rate; 10.8% 75th percentile rate/upper quartile; 4.9% 50th percentile rate/median rate; 2.6% 25th percentile rate/lower quartile; and 0% 10th percentile rate.
CY 2013: 1,4880 denominator cases; 658 numerator cases; 200 hospitals; 4.4% national aggregate rate; 0.05737 mean of hospital rates; 0.10193 standard deviation; 13.9% 90th percentile rate; 7.6%% 75th percentile rate/upper quartile; 2.6% 50th percentile rate/median rate; 0% 25th percentile rate/lower quartile; and 0% 10th percentile rate.
CY 2014: 130,882 denominator cases; 4,331 numerator cases; 1388 hospitals; 3.3% national aggregate rate; 0.03406 mean of hospital rates; 0.04647 standard deviation; 8.7% 90th percentile rate; 4.5% 75th percentile rate/upper quartile; 2.1% 50th percentile rate/median rate; 0% 25th percentile rate/lower quartile; and 0% 10th percentile rate.

1b.4. **Provide disparities data from the measure as specified (current and over time) by population group, e.g., by race/ethnicity, gender, age, insurance status, socioeconomic status, and/or disability.** *(This is required for endorsement maintenance. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities include.) This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*
In a review of two studies by Engle & Kominiarek published in 2008, it was determined that race and ethnicity have an impact on early term deliveries. When comparing Non-Hispanic White births with Hispanic and Black births, they found that the Non-Hispanic White births at 36 weeks accounted for the largest increase in elective cesarean deliveries from 1992 and 2002. This accounted for 3.1% to 3.9% of the total births reviewed in their study.

Interestingly, an overall increase was noted for all three groups. The reason for the increase has not been determined; however, factors speculated to account for the increase include socioeconomic status, access to health care and maternal demand for elective delivery. The rise in induction of labor is present for all racial groups with the highest increase in non-Hispanic whites. Bailey, et al. (2014) recently reviewed data for 638 rural women, recruited prenatally from three counties in rural southern Appalachia, who delivered electively at = 37 weeks. Those delivered electively at early term were 7.7 times more likely to be low birth weight, 4.4 times more likely to have a neonatal intensive care unit admission, and 2.5 times more likely to develop jaundice. Patients living furthest from the hospital were most likely to deliver electively at <39 weeks. Although rates of elective deliveries <39 weeks were no higher than national rates, adjusted odds ratios (aOR) of associated admission to a neonatal intensive care unit doubled (aOR 4.4 vs aOR 2.2).

1b.5. **If no or limited data on disparities from the measure as specified is reported in 1b4, then provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement. Include citations.**
• Bailey, B., McCook, J. & Chaires, C. (2014). Burden of elective early-term births in rural Appalachia. South Med J. 2014 Oct;107(10):624-9.
• Engle, W.A. & Kominiarek, M.A. (2008). Late preterm infants, early term infants, and timing of elective deliveries. Clin Perinatol. 35:325-41.

1c. High Priority (previously referred to as High Impact)

The measure addresses:

- a specific national health goal/priority identified by DHHS or the National Priorities Partnership convened by NQF;
  OR
- a demonstrated high-priority (high-impact) aspect of healthcare (e.g., affects large numbers of patients and/or has a substantial impact for a smaller population; leading cause of morbidity/mortality; high resource use (current and/or future); severity of illness; and severity of patient/societal consequences of poor quality).

1c.1. Demonstrated high priority aspect of healthcare
Frequently performed procedure, Patient/societal consequences of poor quality
1c.2. If Other:

1c.3. Provide epidemiologic or resource use data that demonstrates the measure addresses a high priority aspect of healthcare. List citations in 1c.4.
For almost 3 decades, the American College of Obstetricians and Gynecologists (ACOG) and the American Academy of Pediatrics (AAP) have had in place a standard requiring 39 completed weeks gestation prior to elective delivery, either vaginal or operative (ACOG, 1996). In 2009, ACOG published guidelines listing some of the acceptable medical indications for early induction of labor. Early term deliveries (between 37 and 38 weeks gestation) have increased dramatically from 1990 through 2006 and account for 17.5% of live births in the U.S. (Davidoff et al., 2006). A survey conducted in 2007 of almost 20,000 births in Hospital Corporation of America (HCA) hospitals throughout the U.S. carried out in conjunction with the March of Dimes at the request of ACOG, revealed that almost 1/3 of all babies delivered in the United States (US) are electively delivered earlier than the recommended 39 weeks of gestation (citation). These procedures are conducted without documented evidence supporting medical indication for early delivery. This number of elective cesarean deliveries under 39 weeks gestation, without medical indication, represents 5% of all deliveries in the U.S., those deliveries violating ACOG/AAP guidelines (Clark et al., 2009).

Clark, et.al. (2009) found that most early elective deliveries are for convenience, and result in significant short term neonatal morbidity (neonatal intensive care unit admission rates of 13- 21%). Clark conducted a subsequent retrospective cohort study examining 27 hospitals, and determined that when strategies were implemented to reduce non-medically indicated elective early term deliveries, there was a reduction in elective deliveries of 9.6% to 4.3% (Clark, et. al., 2010). Consequently, the rate of term neonatal intensive care admissions also fell by 16%.

According to Glantz (2005), when comparing spontaneous labor, elective inductions result in more cesarean deliveries and longer maternal length of stay. The American Academy of Family Physicians (2000) also notes that elective induction doubles the cesarean delivery rate. Repeat elective cesarean deliveries before 39 weeks gestation also results in higher rates of adverse respiratory distress syndrome (RDS), mechanical ventilation, sepsis, and hypoglycemia for the newborns (Tita,, et. al., 2009). Newborns born at 37 weeks gestation have a 7.5 fold greater rate of developing RDS versus those born at 39 to 41 weeks gestation (Tita, et al., 2009). Early-term newborns born at 37-38 weeks gestation also are at higher risk for transient tachypnea of the newborn, pulmonary hypertension, hospital stays greater than 5 days as well as diagnoses associated with severe morbidities or death versus newborns delivered at 39 weeks gestation (Engle & Kominiarek, 2008).

Recently, Little et al. (2015) conducted a multi-state analysis of data from 2005 through 2011 and MacDorman et al. (2015) analyzed U.S. data collected by the National Institutes of Health from 2006 through 2012. Both authors concluded that the lack of change in prospective stillbirth rates during these time periods suggests that preventing nonmedically indicated deliveries before 39 weeks of gestation did not increase the U.S. stillbirth rate.

1c.4. Citations for data demonstrating high priority provided in 1a.3
• American Academy of Family Physicians. (2000). Tips from Other Journals: Elective induction doubles cesarean delivery rate, 61, 4.Retrieved September 16, 2011 at: http://www.aafp.org/afp/20000215/tips/39.html.
• American Academy of Pediatrics, American College of Obstetricians and Gynecologists. Guidelines for perinatal care.

7th ed. Elk Grove Village (IL): AAP; Washington, DC: American College of Obstetricians and Gynecologists (2012). p. 109-110, 160, 192-194, 248.
• American College of Obstetricians and Gynecologists (ACOG) Practice Bulletin No 107. (2009). Induction of labor. Obstetrics & Gynecology. 114(2). 386-97.
• American College of Obstetricians and Gynecologists. (2013). Medically indicated late-preterm and early-term deliveries. Committee Opinion No. 560. Am J Obstet Gynecol 2013;121:908–10.
• American College of Obstetricians and Gynecologists. (2013). Nonmedically indicated early-term deliveries. Committee Opinion No. 561. Am J Obstet Gynecol. 121:911–5.
• Clark, S., Miller, D., Belfort, M., Dildy, G., Frye, D., & Meyers, J. (2009). Neonatal and maternal outcomes associated with elective delivery. [Electronic Version]. Am J Obstet Gynecol. 200:156.e1-156.e4.
• Clark, S., Frye, D., Meyers, J., Belfort, M., Dildy, G., Kofford, S et al. (2010). Reduction in elective delivery at <39 weeks of gestation: comparative effectiveness of 3 approaches to change and the impact on neonatal intensive care admission and stillbirth. Am J Obstet Gynecol. 203:449.e1-6.
• Clark, S., Meyers, J., Perlin, J. (2011). Oversight of elective early term deliveries – avoiding unintended consequences, Am J Obstet Gynecol, doi: 10.1016/j.ajog.2011.08.017.
• Clark, S., Meyers, J., Milton, C., Frye, D., Horner, S., Baker, A., & Perlin, J. (2013). Validation of the joint commission exclusion criteria for elective early-term delivery, Am J Obstet Gynecol.0:0. 1-5.
• Davidoff, M., Dias, T., Damus, K., Russell, R., Bettegowda, V.R., Dolan, S., et al. (2006). Changes in the gestational age distribution among U.S. singleton births: impacts on rates of late preterm birth, 1992-2002. Semin Perinatol. Feb; 30(1):8-15.
• Engle, W.A. & Kominiarek, M.A. (2008). Late preterm infants, early term infants, and timing of elective deliveries. Clin Perinatol. 35:325-41.
• Glantz, J. (Apr.2005). Elective induction vs. spontaneous labor associations and outcomes. [Electronic Version]. J Reprod Med. 50(4):235-40.
• Little, S., Zera, C., Clapp, M., Wilkins-Haug, L. & Robinson, J. (2015). A multi-state analysis of early-term delivery trends and the association with term stillbirth. Am J Obstet Gynecol. doi: 10.1097/AOG.0000000000001109.
• MacDorman, M., Reddy, U. & Silver, R. (2015). Trends in stillbirth by gestational age in the united states, 2006-2012. Am J Obstet Gynecol. doi: 10.1097/AOG.0000000000001152.
• Tita, A., Landon, M., Spong, C., Lai, Y., Leveno, K., Varner, M, et al. (2009). Timing of elective repeat cesarean delivery at term and neonatal outcomes. [Electronic Version]. NEJM. 360:2, 111-120.

1c.5. **If a PRO-PM (e.g. HRQoL/functional status, symptom/burden, experience with care, health-related behaviors), provide evidence that the target population values the measured PRO and finds it meaningful.** *(Describe how and from whom their input was obtained.)*
Not Applicable.

2. Reliability and Validity—Scientific Acceptability of Measure Properties

Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. ***Measures must be judged to meet the subcriteria for both reliability and validity to pass this criterion and be evaluated against the remaining criteria.***

2a.1. Specifications The measure is well defined and precisely specified so it can be implemented consistently within and across organizations and allows for comparability. eMeasures should be specified in the Health Quality Measures Format (HQMF) and the Quality Data Model (QDM).

De.5. **Subject/Topic Area** *(check all the areas that apply):*
Perinatal and Reproductive Health, Perinatal and Reproductive Health : Perinatal

De.6. **Cross Cutting Areas** *(check all the areas that apply):*
Overuse, Safety : Complications

S.1. **Measure-specific Web Page** *(Provide a URL link to a web page specific for this measure that contains current detailed specifications including code lists, risk model details, and supplemental materials. Do not enter a URL linking to a home page or to general information.)*
https://www.cms.gov/regulations-and-guidance/legislation/ehrincentiveprograms/ecqm_library.html

S.2a. **If this is an eMeasure**, HQMF specifications must be attached. Attach the zipped output from the eMeasure authoring tool (MAT) - if the MAT was not used, contact staff. (Use the specification fields in this online form for the plain-language description of the specifications)
This is an eMeasure **Attachment:** EH_CMS113v4_NQF0469_PC01.zip

S.2b. **Data Dictionary, Code Table, or Value Sets** *(and risk model codes and coefficients when applicable) must be attached. (Excel or csv file in the suggested format preferred - if not, contact staff)*
Attachment **Attachment:** ElectiveDelivery_v4_Wed_Apr_01_14.49.44_CDT_2015-635908096518042002.xls

S.3. **For endorsement maintenance**, please briefly describe any changes to the measure specifications since last endorsement date and explain the reasons.
On an annual basis, the 29 eCQMs maintained under the H-MDM project and included in the EHR Incentive Program for Eligible Hospitals undergo an annual update to revise specifications based on updated research and clinical information or standards changes. Changes have been made to the eCQM specifications in order to reflect revisions to the chart abstracted measure from which this measure is derived.

S.4. **Numerator Statement** *(Brief, narrative description of the measure focus or what is being measured about the target population, i.e., cases from the target population with the target process, condition, event, or outcome)*
IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
Patients with elective deliveries by either:
- Medical induction of labor while not in labor prior to the procedure
- Cesarean birth while not in labor and with no history of a prior uterine surgery

S.5. **Time Period for Data** *(What is the time period in which data will be aggregated for the measure, e.g., 12 mo, 3 years, look back to August for flu vaccination? Note if there are different time periods for the numerator and denominator.)*
Episode of care

S.6. **Numerator Details** *(All information required to identify and calculate the cases from the target population with the target process, condition, event, or outcome such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required*

format at S.2b)
IF an OUTCOME MEASURE, describe how the observed outcome is identified/counted. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.

- Medical Induction of Labor is represented as a code from one of the following value sets and associated QDM datatype:
  - Procedure, Performed: Medical Induction of Labor (OID 2.16.840.1.113883.3.117.1.7.1.288)
  - Procedure, Performed: Artificial Rupture of Membranes (OID 2.16.840.1.113762.1.4.1045.57)
  - Medication, Administered: Oxytocin (OID 2.16.840.1.113762.1.4.1045.55)
  - Medication, Administered: Dinoprostone (OID 2.16.840.1.113762.1.4.1045.56)

- Labor is represented with the QDM datatype and value set of "Physical Exam, Performed: Labor (OID 2.16.840.1.113883.3.117.1.7.1.281)

- Cesarean Birth is represented with the QDM data type and value set of "Procedure, Performed: Cesarean Birth (OID: 2.16.840.1.113883.3.117.1.7.1.282)

- Prior Uterine Surgery is represented as a code from one of the following value sets and associated QDM datatype:
  - Diagnosis, Resolved: Perforation of Uterus (OID 2.16.840.1.113883.3.117.1.7.1.136)
  - Diagnosis, Resolved: Uterine Window (OID 2.16.840.1.113883.3.117.1.7.1.137)
  - Diagnosis, Resolved: Uterine Rupture (OID 2.16.840.1.113883.3.117.1.7.1.138)
  - Diagnosis, Inactive: Cornual Ectopic Pregnancy (OID 2.16.840.1.113762.1.4.1045.27)
  - Procedure, Performed: Classical Cesarean Birth (OID 2.16.840.1.113883.3.117.1.7.1.421)
  - Procedure, Performed: Myomectomy (OID 2.16.840.1.113883.3.117.1.7.1.422)
  - Procedure, Performed: Transabdominal Cerclage (OID 2.16.840.1.113762.1.4.1110.2)

To access the value sets for the measure, please visit the Value Set Authority Center (VSAC), sponsored by the National Library of Medicine, at this link: https://vsac.nlm.nih.gov/.

S.7. **Denominator Statement** *(Brief, narrative description of the target population being measured)*
The Denominator is patients who deliver newborns with >= 37 and < 39 weeks of gestation completed.

S.8. **Target Population Category** *(Check all the populations for which the measure is specified and tested if any):*
Maternal Health

S.9. **Denominator Details** *(All information required to identify and calculate the target population/denominator such as definitions, specific data collection items/responses , code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*

- Estimated Gestational Age is represented with the QDM datatype and value set of Physical Exam, Performed: Estimated Gestational Age at Delivery (OID: 2.16.840.1.113762.1.4.1045.26)

- Time of Delivery is represented with the QDM datatype and value set of Physical Exam, Performed: Time of Delivery (OID: 2.16.840.1.113762.1.4.1045.28)

S.10. **Denominator Exclusions** *(Brief narrative description of exclusions from the target population)*
ICD-9-CM, ICD-10-CM, or SNOMED CT codes for conditions possibly justifying elective delivery prior to 39 weeks gestation.

S.11. **Denominator Exclusion Details** *(All information required to identify and calculate exclusions from the denominator such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*

- Conditions possibly justifying elective delivery are represented with the QDM datatype and value set Diagnosis, Active: Conditional Possibly Justifying Elective Delivery Prior to 39 Weeks Gestation (OID: 2.16.840.1.113883.3.117.1.7.1.286)

S.12. Stratification Details/Variables *(All information required to stratify the measure results including the stratification variables, definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format with at S.2b)*
Not Applicable, the measure is not stratified

S.13. Risk Adjustment Type (Select type. Provide specifications for risk stratification in S.12 and for statistical model in S.14-15)
No risk adjustment or risk stratification
If other:

S.14. Identify the statistical risk model method and variables *(Name the statistical method - e.g., logistic regression and list all the risk factor variables. Note - risk model development and testing should be addressed with measure testing under Scientific Acceptability)*
Not Applicable

S.15. Detailed risk model specifications *(must be in attached data dictionary/code list Excel or csv file. Also indicate if available at measure-specific URL identified in S.1.)*
Note: Risk model details (including coefficients, equations, codes with descriptors, definitions), should be provided on a separate worksheet in the suggested format in the Excel or csv file with data dictionary/code lists at S.2b.

S.15a. Detailed risk model specifications (*if not provided in excel or csv file at S.2b*)
Not Applicable

S.16. Type of score:
Rate/proportion
If other:

S.17. Interpretation of Score *(Classifies interpretation of score according to whether better quality is associated with a higher score, a lower score, a score falling within a defined interval, or a passing score)*
Better quality = Lower score

S.18. Calculation Algorithm/Measure Logic *(Describe the calculation of the measure score as an ordered sequence of steps including identifying the target population; exclusions; cases meeting the target process, condition, event, or outcome; aggregating data; risk adjustment; etc.)*
See attached HQMF file.

S.19. Calculation Algorithm/Measure Logic Diagram URL or Attachment *(You also may provide a diagram of the Calculation Algorithm/Measure Logic described above at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
Available at measure-specific web page URL identified in S.1

S.20. Sampling *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*
IF a PRO-PM, identify whether (and how) proxy responses are allowed.
Not Applicable.

S.21. Survey/Patient-reported data *(If measure is based on a survey, provide instructions for conducting the survey and guidance on minimum response rate.)*

IF a PRO-PM, specify calculation of response rates to be reported with performance measure results.
Not Applicable. This measure is not based on a survey or a PRO-PM

S.22. **Missing data** (specify how missing data are handled, e.g., imputation, delete case.)
Required for Composites and PRO-PMs.
eMeasures are calculated using only the structured data collected in certified EHR technology (CEHRT). Data not present in the structured field from which the measure draws will not be included in the measure calculation.

S.23. **Data Source** *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*
If other, please describe in S.24.
Electronic Clinical Data, Electronic Clinical Data : Electronic Health Record, Electronic Clinical Data : Pharmacy

S.24. **Data Source or Collection Instrument** *(Identify the specific data source/data collection instrument e.g. name of database, clinical registry, collection instrument, etc.)*
IF a PRO-PM, identify the specific PROM(s); and standard methods, modes, and languages of administration.
Hospitals report EHR data using Certified Electronic Health Record Technology (CEHRT), and by submitting Quality Reporting Document Architecture Category 1 (QRDA-1).

S.25. **Data Source or Collection Instrument** *(available at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
No data collection instrument provided

S.26. **Level of Analysis** *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*
Facility, Population : National

S.27. **Care Setting** *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*
Hospital/Acute Care Facility
If other:

S.28. **COMPOSITE Performance Measure** - Additional Specifications *(Use this section as needed for aggregation and weighting rules, or calculation of individual performance measures if not individually endorsed.)*
Not Applicable

2a. Reliability – See attached Measure Testing Submission Form
2b. Validity – See attached Measure Testing Submission Form
PC-01_eCQM_testing_attachment.docx,NQF2829_CMS113v3_PC01_Bonnie_Testing-635908104856946078.xlsx

Measure Number (*if previously endorsed*)**:** 2829

Measure Title: PC-01: Elective Delivery

Date of Submission: 2/11/2016

Type of Measure:

☐ Composite – ***STOP – use composite testing form***	☐ Outcome (*including PRO-PM*)
☐ Cost/resource	X Process
☐ Efficiency	☐ Structure

Instructions

- Measures must be tested for all the data sources and levels of analyses that are specified. ***If there is more than one set of data specifications or more than one level of analysis, contact NQF staff*** about how to present all the testing information in one form.
- **For all measures, sections 1, 2a2, 2b2, 2b3, and 2b5 must be completed.**
- **For outcome and resource use measures**, section **2b4** also must be completed.
- If specified for **multiple data sources/sets of specificaitons** (e.g., claims and EHRs), section **2b6** also must be completed.
- Respond to all questions as instructed with answers immediately following the question. All information on testing to demonstrate meeting the sub criteria for reliability (2a2) and validity (2b2-2b6) must be in this form. An appendix for *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 20 pages (*incuding questions/instructions;* minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.
- For information on the most updated guidance on how to address sociodemographic variables and testing in this form refer to the release notes for version 6.6 of the Measure Testing Attachment.

Note: The information provided in this form is intended to aid the Steering Committee and other stakeholders in understanding to what degree the testing results for this measure meet NQF's evaluation criteria for testing.

2a2. Reliability testing [10] demonstrates the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise. For **PRO-PMs and composite performance measures**, reliability should be demonstrated for the computed performance score.

2b2. Validity testing [11] demonstrates that the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality. For **PRO-PMs and composite performance measures**, validity should be demonstrated for the computed performance score.

2b3. Exclusions are supported by the clinical evidence; otherwise, they are supported by evidence of sufficient frequency of occurrence so that results are distorted without the exclusion; [12]
AND
If patient preference (e.g., informed decision-making) is a basis for exclusion, there must be evidence that the exclusion impacts performance on the measure; in such cases, the measure must be specified so that the information about patient preference and the effect on the measure is transparent (e.g., numerator category computed separately, denominator exclusion category computed separately). [13]

2b4. For outcome measures and other measures when indicated (e.g., resource use):

- **an evidence-based risk-adjustment strategy** (e.g., risk models, risk stratification) is specified; is based on patient factors (including clinical and sociodemographic factors) that influence the measured outcome and are present at start of care; [14],[15] and has demonstrated adequate discrimination and calibration

OR

- rationale/data support no risk adjustment/ stratification.

2b5. Data analysis of computed measure scores demonstrates that methods for scoring and analysis of the specified measure allow for **identification of statistically significant and practically/clinically meaningful [16] differences in performance**;

OR

there is evidence of overall less-than-optimal performance.

2b6. If multiple data sources/methods are specified, there is demonstration they produce comparable results.

2b7. For **eMeasures, composites, and PRO-PMs** (or other measures susceptible to missing data), analyses identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and non-responders) and how the specified handling of missing data minimizes bias.

Notes

10. Reliability testing applies to both the data elements and computed measure score. Examples of reliability testing for data elements include, but are not limited to: inter-rater/abstractor or intra-rater/abstractor studies; internal consistency for multi-item scales; test-retest for survey items. Reliability testing of the measure score addresses precision of measurement (e.g., signal-to-noise).

11. Validity testing applies to both the data elements and computed measure score. Validity testing of data elements typically analyzes agreement with another authoritative source of the same information. Examples of validity testing of the measure score include, but are not limited to: testing hypotheses that the measures scores indicate quality of care, e.g., measure scores are different for groups known to have differences in quality assessed by another valid quality measure or method; correlation of measure scores with another valid indicator of quality for the specific topic; or relationship to conceptually related measures (e.g., scores on process measures to scores on outcome measures). Face validity of the measure score as a quality indicator may be adequate if accomplished through a systematic and transparent process, by identified experts, and explicitly addresses whether performance scores resulting from the measure as specified can be used to distinguish good from poor quality.

12. Examples of evidence that an exclusion distorts measure results include, but are not limited to: frequency of occurrence, variability of exclusions across providers, and sensitivity analyses with and without the exclusion.

13. Patient preference is not a clinical exception to eligibility and can be influenced by provider interventions.

14. Risk factors that influence outcomes should not be specified as exclusions

15. With large enough sample sizes, small differences that are statistically significant may or may not be practically or clinically meaningful. The substantive question may be, for example, whether a statistically significant difference of one percentage point in the percentage of patients who received smoking cessation counseling (e.g., 74 percent v. 75 percent) is clinically meaningful; or whether a statistically significant difference of $25 in cost for an episode of care (e.g., $5,000 v. $5,025) is practically meaningful. Measures with overall less-than-optimal performance may not demonstrate much variability across providers.

1. DATA/SAMPLE USED FOR ALL TESTING OF THIS MEASURE

Often the same data are used for all aspects of measure testing. In an effort to eliminate duplication, the first five questions apply to all measure testing. If there are differences by aspect of testing,(e.g., reliability vs. validity) be sure to indicate the specific differences in question 1.7.

1.1. What type of data was used for testing? (*Check all the sources of data identified in the measure specifications and data used for testing the measure*. *Testing must be provided for all the sources of data specified and intended for measure implementation.* ***If different data sources are used for the numerator and denominator, indicate N [numerator] or D [denominator] after the checkbox.***)

Measure Specified to Use Data From:	Measure Tested with Data From:

(*must be consistent with data sources entered in S.23*)	
☐ abstracted from paper record	☐ abstracted from paper record
☐ administrative claims	☐ administrative claims
☐ clinical database/registry	☐ clinical database/registry
☐ abstracted from electronic health record	☐ abstracted from electronic health record
☒ eMeasure (HQMF) implemented in EHRs	☐ eMeasure (HQMF) implemented in EHRs
☒other: Bonnie Test Cases	☒ other: Bonnie Test Cases (see question 1.2)

1.2. If an existing dataset was used, identify the specific dataset (*the dataset used for testing must be consistent with the measure specifications for target population and healthcare entities being measured; e.g., Medicare Part A claims, Medicaid claims, other commercial insurance, nursing home MDS, home health OASIS, clinical registry*).

This measure is a legacy eCQM (an NQF endorsed measure that has been re-specified into an eMeasure and is currently used in a federal quality program/programs). In accordance with the modified NQF requirements for testing of eMeasures, the Bonnie testing tool was used to simulate a testing environment where measure specifications and HQMF output are tested against synthetic test data. Measure developers rely on the results in Bonnie to confirm whether the measure logic is performing as expected. Reference the eCQI Resource Center website (https://ecqi.healthit.gov/ecqm-tools/tool-library/bonnie) or the Bonnie testing tool website (https://bonnie.healthit.gov/) for more information about Bonnie functionality and its role in measure development. Please also reference the Bonnie testing worksheet attachment for detailed Bonnie test cases and testing results for this measure.

While Bonnie does consume HQMF specifications, we are not able to test the measure with data from HQMF implemented in EHRs.

1.3. What are the dates of the data used in testing? The measure specifications were tested in the Bonnie testing environment which mimics the year 2012.

1.4. What levels of analysis were tested? (*testing must be provided for all the levels specified and intended for measure implementation, e.g., individual clinician, hospital, health plan*)

Measure Specified to Measure Performance of: (***must be consistent with levels entered in item S.26***)	**Measure Tested at Level of:**
☐ individual clinician	☐ individual clinician
☐ group/practice	☐ group/practice
☒ hospital/facility/agency	☒ hospital/facility/agency
☐ health plan	☐ health plan
☐ other:	☒ other: Bonnie Test Cases

1.5. How many and which measured entities were included in the testing and analysis (by level of analysis and data source)? (*Identify the number and descriptive characteristics of measured entities included in the analysis (e.g., size, location, type); if a sample was used, describe how entities were selected for inclusion in the sample*)

Not applicable.

1.6. How many and which patients were included in the testing and analysis (by level of analysis and data source)? (*Identify the number and descriptive characteristics of patients included in the analysis (e.g., age, sex, race, diagnosis); if a sample was used, describe how patients were selected for inclusion in the sample*)

51 unique synthetic patient records were created in the BONNIE testing system for this measure. Cases were used to test the validity of each data element and timing relationship in the measure. Patient characteristics such as age, diagnosis, and length of stay were pre-determined to provide a variety of scenarios that

adequately tested for patients passing each data element and failing each data element. Data included in cases and tested for this measure included diagnoses, differing types of delivery and uterine procedures, gestational ages, medications, patient orders, age, length of stay, and clinical observations, such as time of delivery.

All 51 cases passed or failed as expected based on the data included in the case, confirming the measure logic is accurate and valid.

1.7. If there are differences in the data or sample used for different aspects of testing (e.g., reliability, validity, exclusions, risk adjustment), identify how the data or sample are different for each aspect of testing reported below.

Not applicable. Bonnie test patients were developed for use across the different aspects of testing.

1.8 What were the patient-level sociodemographic (SDS) variables that were available and analyzed in the data or sample used? For example, patient-reported data (e.g., income, education, language), proxy variables when SDS data are not collected from each patient (e.g. census tract), or patient community characteristics (e.g. percent vacant housing, crime rate).

Not applicable.

2a2. RELIABILITY TESTING

Note*: If accuracy/correctness (validity) of data elements was empirically tested, separate reliability testing of data elements is not required – in 2a2.1 check critical data elements; in 2a2.2 enter "see section 2b2 for validity testing of data elements"; and skip 2a2.3 and 2a2.4.*

2a2.1. What level of reliability testing was conducted? (*May be one or both levels*)

X Critical data elements used in the measure (*e.g., inter-abstractor reliability; data element reliability must address ALL critical data elements*)

☐ **Performance measure score** (e.g., *signal-to-noise analysis*)

2a2.2. For each level checked above, describe the method of reliability testing and what it tests (*describe the steps—do not just name a method; what type of error does it test; what statistical analysis was used*)

The chart-abstracted version of this measure has been in national use since the 2nd quarter of 2010. At the time the chart-abstracted measure was originally tested, extensive tests of measure reliability were conducted. At present, no performance data for the electronic version of the measure are currently available. Below is the reliability testing summary for NQF #0469 PC-01 Elective Delivery Prior to 39 Completed Weeks Gestation, from which this measure is derived.

Demographics of organizations collecting and reporting data on these measures are as follows:
163 health care organizations representing various types, locations and sizes:
10 For Profit, 91 Not for Profit, 46 Military Facilities, 9 County, 2 State, 5 Other
15 >=500 beds; 29 250-499 beds; 50 100-249 beds; 69 <100 beds
Located in: AE, AK, AL, AP, AR, AZ, CA, DO, DC, FL, GA, HI, ID, IL, IN, KS, KY, LA, MA, MD, MI, MN, MO, MS, MT, NC, NE, NV, NY, OH,OK, PA, PR, RI, SC, TN, TX, VA, WA, WI, WV
26 performance measurement systems

Currently, hospitals are supported in their data collection and reporting efforts by 26 contracted performance measurement system (PMS) vendors. It is a contractual requirement of Joint Commission listed vendors that the quality and reliability of data submitted to them by contracted health care organizations must be monitored on a quarterly

basis. In addition, The Joint Commission analyzes these data by running 17 quality tests on the data submitted into ORYX. (ORYX is the term used by The Joint Commission to describe the component of the hospital accreditation program which requires data collection and reporting on standardized national performance measures). The following is a list of the major tests done on the submitted ORYX data, taken from the 2011 ORYX Performance Measurement System Requirements manual.

- Transmission of complete data
- Usage of individual core measure data received: To understand if the HCO provides the relevant service to treat the relevant population
- Investigation of aberrant data points
- Verification of patient population and sample size
- Identification of missing data elements
- Validation of the accuracy of target outliers
- Data integrity
- Data corrections

Data Element Agreement Rate:
Inter-rater reliability testing methodology utilized by contracted performance measure system vendors as outlined in the contract is as follows:

- All clinical data elements and all editable demographic elements are scored.
- All measure data are reabstracted with originally abstracted data having been blinded so that the reabstraction is not biased.
- Reabstracted data are compared with originally abstracted data on a data element by data element basis. A data element agreement rate is calculated. Clinical and demographic data are scored separately, and an overall agreement rate is computed.

2a2.3. For each level of testing checked above, what were the statistical results from reliability testing? (e.*g., percent agreement and kappa for the critical data elements; distribution of reliability statistics from a signal-to-noise analysis*)

Data element agreement rates for the chart-abstracted version of this measure were reported to The Joint Commission for 1Q11. This reflects the findings of 108 hospitals, comprising 13,279 records (100% sample). The following table delineates calculated agreement rates for individual data elements that are used to compute measure rates for PC-01.

Data Elements with a Mismatch - Mother	total n	total d	rate
Active Labor	33	35	94.29%
Gestational Age	639	712	89.75%

These agreement rates are considered to be well within acceptable levels

2a2.4 What is your interpretation of the results in terms of demonstrating reliability? (i.*e., what do the results mean and what are the norms for the test conducted?*)

Agreement rates for individual data elements tested for the chart-abstracted version of this measure were within acceptable levels. Once data are available for analysis, it is expected that reliability tests of the eCQM version of this measure will yield similar results.

2b2. VALIDITY TESTING

2b2.1. What level of validity testing was conducted? (*may be one or both levels*)

- ☒ **Critical data elements** (*data element validity must address ALL critical data elements*)
- ☒ **Performance measure score**

☐ **Empirical validity testing**

☒ **Systematic assessment of face validity of performance measure score as an indicator** of quality or resource use (*i.e., is an accurate reflection of performance on quality or resource use and can distinguish good from poor performance*)

2b2.2. For each level of testing checked above, describe the method of validity testing and what it tests (*describe the steps—do not just name a method; what was tested, e.g., accuracy of data elements compared to authoritative source, relationship to another measure as expected; what statistical analysis was used)*

The Bonnie testing tool and environment were used to test the measure logic and value sets. Each data element and logic statement was tested to confirm actual results met expectations. Bonnie testing includes negative and positive testing of each data element in the measure. Positive testing ensures patients expected to be included in the measure are included. Negative testing ensures that patients who do not meet the data criteria are not included in the measure. An example of negative testing would be to include test cases with pediatric ages to ensure that pediatric patients are not included in the measure.

Denominator test cases positively test to ensure that only patients >=8 and <65 years of age who deliver an infant >=37 weeks and <39 weeks at time of delivery are included in the denominator. Negative test cases ensure that patients who do not meet these criteria to do not pass into the denominator. For example, cases test patient ages of 7, 8, 64, and 65 to ensure that patient age 7 and 65 do not pass into the denominator, but patients age 8 and 64 are included.

Numerator test cases positively test to ensure patients with elective deliveries by either medical induction of labor prior to the start of labor, or by cesarean birth while not in labor and with no history of prior uterine surgery are included in the numerator. Negative test cases ensure that a patient who did not meet these criteria are not included. For example, test cases in which labor began prior to cesarean birth ensure that patients who have a cesarean procedure as a result of complications of labor are not included in the numerator.

Denominator exclusion test cases for this measure ensure that patients are properly removed from the denominator if they have documented diagnoses that reflect conditions possibly justifying elective delivery. Negative test cases for the denominator exclusion ensure that patients without these diagnoses fall in to the denominator population. Testing confirmed patients meeting the exclusion and exception criteria are removed from the measure appropriately, while those that do not meet the criteria are retained in the denominator population.

A review of the measure specifications was also conducted to confirm the logic was properly expressed within the current version of the QDM and confirmed the logic matches the clinical intent of the measure, as stated in the measure header. This is done through use of a logic checklist to facilitate review of the measure logic, according to several checks, a few examples of which are included below:

- Is the intent of the measure described in the measure description articulated/ captured in the measure logic?
- Do the logic elements map to definitions in the measure narrative, data dictionary or supporting reference documentation?
- Do the populations in the narrative align with the populations defined in the logic?
- Are the mathematic inequalities reflective of the measure intent and represent the intended populations (for example: when intended the inequality represents less than rather than less and or equal to)?

2b2.3. What were the statistical results from validity testing? (*e.g., correlation; t-test*)

Bonnie results provide coverage and passing rates. This measure reached 100% coverage, confirming there is a test case for each pathway of logic. This measure also has a 100% passing rate, confirming all test cases performed as expected.

2b2.4. What is your interpretation of the results in terms of demonstrating validity? (i.*e., what do the results mean and what are the norms for the test conducted?*)

This measure performed as expected in Bonnie.

_____

2b3. EXCLUSIONS ANALYSIS

NA ☐ no exclusions — *skip to section 2b4*

2b3.1. Describe the method of testing exclusions and what it tests (*describe the steps—do not just name a method; what was tested, e.g., whether exclusions affect overall performance scores; what statistical analysis was used*)

Exclusions in the eCQM align with the chart-based version of the measure, and are clinically necessary for the interpretation of the measure. As noted previously, the chart-abstracted version of this measure has been in national use since the 2nd quarter of 2010, and no data are available for the eCQM version of the measure. The below analysis addresses exclusions testing performed for the chart-abstracted version of the measure from which this measure is derived.

There were 1,134,640 admissions selected from the initial cohort. From among the 1,134,640 admissions in 1,237 hospitals, the descriptive statistics are given below.

The following exclusions were analyzed by subpopulation and measure for frequency and variability across providers:

- ICD-9-CM Principal Diagnosis Code or ICD-9-CM Other Diagnosis Codes for conditions possibly justifying elective delivery prior to 39 weeks gestation as defined in Appendix A, Table 11.07
- Less than 8 years of age
- Greater than or equal to 65 years of age
- Length of stay > 120 days
- Enrolled in clinical trials
- Gestational Age < 37 or >= 39 weeks or UTD

2b3.2. What were the statistical results from testing exclusions? (*include overall number and percentage of individuals excluded, frequency distribution of exclusions across measured entities, and impact on performance measure scores*)

Exclusion Subpopulation 1 - PC-01

ICD-9-CM Principal Diagnosis Code or ICD-9-CM Other Diagnosis Codes for conditions possibly justifying elective delivery prior to 39 weeks gestation as defined in Appendix A, Table 11.07

Exclusion: No observations noted

Less than 8 years of age

Exclusion: Included in the initial population exclusion

Greater than or equal to 65 years of age

Exclusion: Included in the initial population exclusion

Length of Stay >120 days

Exclusion: Included in the initial population exclusion

Exclusion: Enrolled in Clinical Trials

Overall Number of Occurrences n = 748

Overall Occurrence Percentage: 0.07%

Minimum: 0%

10th Percentile: 0%

Median: 0%
90th Percentile: 0.062%
Maximum: 28%

Exclusion: Gestational Age < 37 or gestational Age = >39 weeks or UTD
Overall Number of Occurrences n = 851,258
Overall Occurrence Percentage: 84.9%
Minimum 0.29%
10th Percentile: 69.17%
Median: 75.2%
90th Percentile: 79.2%
Maximum: 84.8%

2b3.3. What is your interpretation of the results in terms of demonstrating that exclusions are needed to prevent unfair distortion of performance results? (*i.e., the value outweighs the burden of increased data collection and analysis. Note: **If patient preference is an exclusion**, the measure must be specified so that the effect on the performance score is transparent, e.g., scores with and without exclusion*)

Analysis of these data for the chart-abstracted progenitor of this measure indicated that all exclusions were appropriate. It is believed that results for exclusions in the eCQM will be similar when sufficient data have been received to perform such an analysis.

2b4. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES
If not an intermediate or health outcome, or PRO-PM, or resource use measure, skip to section 2b5.

2b4.1. What method of controlling for differences in case mix is used?

- ☒ **No risk adjustment or stratification**
- ☐ **Statistical risk model with** Click here to enter number of factors **risk factors**
- ☐ **Stratification by** Click here to enter number of categories **risk categories**
- ☐ **Other,** Click here to enter description

2b4.2. If an outcome or resource use measure is not risk adjusted or stratified, provide rationale and analyses to demonstrate that controlling for differences in patient characteristics (case mix) is not needed to achieve fair comparisons across measured entities.

2b4.3. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors (clinical factors or sociodemographic factors) used in the statistical risk model or for stratification by risk (*e.g., potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of p<0.10; correlation of x or higher; patient factors should be present at the start of care*)

2b4.4a. What were the statistical results of the analyses used to select risk factors?

2b4.4b. Describe the analyses and interpretation resulting in the decision to select SDS factors (e.g. prevalence of the factor across measured entities, empirical association with the outcome, contribution of unique variation in the outcome, assessment of between-unit effects and within-unit effects)

2b4.5. Describe the method of testing/analysis used to develop and validate the adequacy of the statistical model or stratification approach (*describe the steps—do not just name a method; what statistical analysis was used*)

Provide the statistical results from testing the approach to controlling for differences in patient characteristics (case mix) below.
If stratified, skip to 2b4.9

2b4.6. Statistical Risk Model Discrimination Statistics (*e.g., c-statistic, R-squared*)**:**

2b4.7. Statistical Risk Model Calibration Statistics (*e.g., Hosmer-Lemeshow statistic*):

2b4.8. Statistical Risk Model Calibration – Risk decile plots or calibration curves:

2b4.9. Results of Risk Stratification Analysis:

2b4.10. What is your interpretation of the results in terms of demonstrating adequacy of controlling for differences in patient characteristics (case mix)? (i*.e., what do the results mean and what are the norms for the test conducted*)

2b4.11. Optional Additional Testing for Risk Adjustment (*not required, but would provide additional support of adequacy of risk model, e.g., testing of risk model in another data set; sensitivity analysis for missing data; other methods that were assessed*)

_____

2b5. IDENTIFICATION OF STATISTICALLY SIGNIFICANT & MEANINGFUL DIFFERENCES IN PERFORMANCE

2b5.1. Describe the method for determining if statistically significant and clinically/practically meaningful differences in performance measure scores among the measured entities can be identified (*describe the steps—do not just name a method; what statistical analysis was used? Do not just repeat the information provided related to performance gap in 1b)*

The method used to analyze meaningful differences in performance at The Joint Commission is Target Analysis. The object of target analysis is to compare a health care organization's (HCO) data against a comparative norm for the purpose of evaluating performance improvement opportunities. When an organization's performance level is statistically significantly different from a comparative norm, it is considered a statistical deviation. A statistical deviation may be desirable or undesirable depending on the "direction of improvement" of the measure.

There are two components to the target analysis methodology used at The Joint Commission. Given the national average for a performance measure, a target range is constructed. Using generalized linear mixed models methodology (also known as hierarchical models), a predicted estimate of an HCO's performance, with a corresponding 95% confidence interval, is generated. This confidence interval is compared to the target range, to determine the HCOs' rating. The estimate of the organization's true performance is based on both the data from that organization and on data from the entire set of reporting organizations. A similar methodology will be used for the eCQMs.

2b5.2. What were the statistical results from testing the ability to identify statistically significant and/or clinically/practically meaningful differences in performance measure scores across measured entities? (e.g., *number and percentage of entities with scores that were statistically significantly different from mean or some benchmark, different from expected; how was meaningful difference defined*)

NQF: 0469: PC-01 Distribution of Outliers

2011 1st Quarter Data:
Scores on this measure: N=160, Mean 13.6%, SD 0.1594

10th Percentile= 0%
25th Percentile= 0%
50th Percentile= 9%
75th Percentile= 19%
90th Percentile= 34%

156 (97.5%) Neutral – results not significantly different from target range
4 (2.5%) Unfavorable - results statistically significantly lower than the national rate

2b5.3. What is your interpretation of the results in terms of demonstrating the ability to identify statistically significant and/or clinically/practically meaningful differences in performance across measured entities? (i.*e., what do the results mean in terms of statistical and meaningful differences?*)

It should be noted that since data collection on this measure is completely voluntary for The Joint Commission, the hospitals reporting on this measure are self-selected, and therefore, presumably have a particular interest and concern for improving perinatal care. For this reason, the measure results demonstrated by this group most likely significantly overstates the rate for the population of all health care organizations

---

2b6. COMPARABILITY OF PERFORMANCE SCORES WHEN MORE THAN ONE SET OF SPECIFICATIONS
If only one set of specifications, this section can be skipped.

Note: *This item is directed to measures that are risk-adjusted (with or without SDS factors)* ***OR*** *to measures with more than one set of specifications/instructions (e.g., one set of specifications for how to identify and compute the measure from medical record abstraction and a different set of specifications for claims or eMeasures). It does not apply to measures that use more than one source of data in one set of specifications/instructions (e.g., claims data to identify the denominator and medical record abstraction for the numerator).* ***Comparability is not required when comparing performance scores with and without SDS factors in the risk adjustment model. However, if comparability is not demonstrated for measures with more than one set of specifications/instructions, the different specifications (e.g., for medical records vs. claims) should be submitted as separate measures.***

2b6.1. Describe the method of testing conducted to compare performance scores for the same entities across the different data sources/specifications (*describe the steps—do not just name a method; what statistical analysis was used*)

Multiple data sources are not used for this measure.

2b6.2. What were the statistical results from testing comparability of performance scores for the same entities when using different data sources/specifications? (*e.g., correlation, rank order*)

Not applicable.

2b6.3. What is your interpretation of the results in terms of the differences in performance measure scores for the same entities across the different data sources/specifications? (i.*e., what do the results mean and what are the norms for the test conducted*)

Not applicable.

2b7. MISSING DATA ANALYSIS AND MINIMIZING BIAS

2b7.1. Describe the method of testing conducted to identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias (*describe the steps—do not just name a method; what statistical analysis was used*)

Data not present in the structured field from which the measure draws will not be included in the measure calculation. In the Bonnie testing environment, missing data are tested as an expected "Fail" for that data element, and actual performance (whether or not the case fails) is compared to expectations to ensure missing data impacts data element and overall measure calculation as expected. This is the extent of missing data analysis that can be performed in Bonnie.

2b7.2. What is the overall frequency of missing data, the distribution of missing data across providers, and the results from testing related to missing data? (*e.g., results of sensitivity analysis of the effect of various rules for missing data/nonresponse; if no empirical sensitivity analysis, identify the approaches for handling missing data that were considered and pros and cons of each*)

See response for question 2b7.1 above.

2b7.3. What is your interpretation of the results in terms of demonstrating that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias**?** (i.*e., what do the results mean in terms of supporting the selected approach for missing data and what are the norms for the test conducted; if no empirical analysis, provide rationale for the selected approach for missing data*)

The measure performed as expected in the Bonnie testing environment.

3. Feasibility

Extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

3a. Byproduct of Care Processes

For clinical measures, the required data elements are routinely generated and used during care delivery (e.g., blood pressure, lab test, diagnosis, medication order).

3a.1. **Data Elements Generated as Byproduct of Care Processes.**

Generated or collected by and used by healthcare personnel during the provision of care (e.g., blood pressure, lab value, diagnosis, depression score)

If other:

3b. Electronic Sources

The required data elements are available in electronic health records or other electronic sources. If the required data are not in electronic health records or existing electronic sources, a credible, near-term path to electronic collection is specified.

3b.1. **To what extent are the specified data elements available electronically in defined fields?** *(i.e., data elements that are needed to compute the performance measure score are in defined, computer-readable fields)*

ALL data elements are in defined fields in electronic health records (EHRs)

3b.2. **If ALL the data elements needed to compute the performance measure score are not from electronic sources, specify a credible, near-term path to electronic capture, OR provide a rationale for using other than electronic sources.**

3b.3. **If this is an eMeasure, provide a summary of the feasibility assessment in an attached file or make available at a measure-specific URL.**

Attachment **Attachment:** PC01_eCQM_NQF_Measure_Feasibility_Assessment_Report-635908844084639084.docx

3c. Data Collection Strategy

Demonstration that the data collection strategy (e.g., source, timing, frequency, sampling, patient confidentiality, costs associated with fees/licensing of proprietary measures) can be implemented (e.g., already in operational use, or testing demonstrates that it is ready to put into operational use). For eMeasures, a feasibility assessment addresses the data elements and measure logic and demonstrates the eMeasure can be implemented or feasibility concerns can be adequately addressed.

3c.1. **Describe what you have learned/modified as a result of testing and/or operational use of the measure regarding data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, time and cost of data collection, other feasibility/implementation issues.**
IF a PRO-PM, consider implications for both individuals providing PROM data (patients, service recipients, respondents) and those whose performance is being measured.

Not Applicable.

3c.2. **Describe any fees, licensing, or other requirements to use any aspect of the measure as specified** *(e.g., value/code set, risk model, programming code, algorithm)*.

Value sets are housed in the Value Set Authority Center (VSAC), which is provided by the National Library of Medicine (NLM), in coordination with the Office of the National Coordinator for Health Information Technology and the Centers for Medicare & Medicaid Services.

Viewing or downloading value sets requires a free Unified Medical Language System® (UMLS) Metathesaurus License, due to usage restrictions on some of the codes included in the value sets. Individuals interested in accessing value set

content can request a UMLS license at (https://uts.nlm.nih.gov/license.html)

There are no other fees or licensing requirements to use the Joint Commission performance measures, all of which are in the public domain.

4. Usability and Use

Extent to which potential audiences (e.g., consumers, purchasers, providers, policy makers) are using or could use performance results for both accountability and performance improvement to achieve the goal of high-quality, efficient healthcare for individuals or populations.

4a. Accountability and Transparency

Performance results are used in at least one accountability application within three years after initial endorsement and are publicly reported within six years after initial endorsement (or the data on performance results are available). If not in use at the time of initial endorsement, then a credible plan for implementation within the specified timeframes is provided.

4.1. Current and Planned Use

NQF-endorsed measures are expected to be used in at least one accountability application within 3 years and publicly reported within 6 years of initial endorsement in addition to performance improvement.

Planned	Current Use (for current use provide URL)
Public Reporting	Payment Program Hospital Inpatient Quality Reporting Program https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/HospitalQualityInits/HospitalRHQDAPU.html Regulatory and Accreditation Programs Hospital Accreditation Program http://jointcommission.org EHR Incentive Program https://www.cms.gov/Regulations-and-Guidance/Legislation/EHRIncentivePrograms/index.html?redirect=/ehrincentiveprograms

4a.1. For each CURRENT use, checked above, provide:

- Name of program and sponsor
- Purpose
- Geographic area and number and percentage of accountable entities and patients included

- Name of program and sponsor: Hospital Inpatient Quality Reporting Program; Centers for Medicare & Medicaid Services
- Purpose: The Hospital Inpatient Quality Reporting (Hospital IQR) program was originally mandated by Section 501(b) of the Medicare Prescription Drug, Improvement, and Modernization Act (MMA) of 2003. This section of the MMA authorized CMS to pay hospitals that successfully report designated quality measures a higher annual update to their payment rates.
- Geographic area and number and percentage of accountable entities and patients included: Nationwide; 3500 hospitals (2015)

- Name of program and sponsor: EHR Incentive Program; Centers for Medicare & Medicaid Services
- Purpose: The American Recovery and Reinvestment Act of 2009 (ARRA) (Pub.L. 111–5) was enacted on February 17, 2009. Title IV of Division B of ARRA amends Titles XVIII and XIX of the Social Security Act (the Act) by establishing incentive payments to eligible professionals (EPs), eligible hospitals, and critical access hospitals (CAHs), and Medicare Advantage Organizations to promote the adoption and meaningful use of interoperable health information technology (HIT) and qualified electronic health records (EHRs). These incentive payments are part of a broader effort under the HITECH Act to accelerate the adoption of HIT and utilization of qualified EHRs.
- Geographic area and number and percentage of accountable entities and patients included: Nationwide; 4,847

active registrations (September, 2015)

- Name of program and sponsor Hospital Accreditation Program; The Joint Commission
- Purpose: An accreditation program that recognizes hospitals that meet standard requirements to provide safe and effective patient care.
- Geographic area and number and percentage of accountable entities and patients included Nationwide; 3300 Joint Commission-accredited hospitals (2014)

4a.2. **If not currently publicly reported OR used in at least one other accountability application (e.g., payment program, certification, licensing) what are the reasons?** (*e.g., Do policies or actions of the developer/steward or accountable entities restrict access to performance results or impede implementation?*)
Not Applicable.

4a.3. **If not currently publicly reported OR used in at least one other accountability application, provide a credible plan for implementation within the expected timeframes -- any accountability application within 3 years and publicly reported within 6 years of initial endorsement.** (*Credible plan includes the specific program, purpose, intended audience, and timeline for implementing the measure within the specified timeframes. A plan for accountability applications addresses mechanisms for data aggregation and reporting.*)
Not Applicable.

4b. Improvement

Progress toward achieving the goal of high-quality, efficient healthcare for individuals or populations is demonstrated. If not in use for performance improvement at the time of initial endorsement, then a credible rationale describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.

4b.1. **Progress on Improvement. (Not required for initial endorsement unless available.)**
Performance results on this measure (current and over time) should be provided in 1b.2 and 1b.4. Discuss:

- **Progress (trends in performance results, number and percentage of people receiving high-quality healthcare)**
- **Geographic area and number and percentage of accountable entities and patients included**

While the chart-abstracted version of this measure has indicated improvement over time, this measure is a legacy eCQM that is currently in use in the Hospital Inpatient Quality Reporting Program (HIQR) and the EHR Incentive Program. At present, no performance data are available.

In CY2016, CMS is requiring organizations participating in HIQR to electronically submit 1 quarter of data on 4 of 28 available eCQMs, one of which is this measure. Data will be accepted through February 28th, 2017.

For more information, refer to CY2016 IPPS Final Rule, located here: https://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/AcuteInpatientPPS/FY2016-IPPS-Final-Rule-Home-Page.html

4b.2. **If no improvement was demonstrated, what are the reasons? If not in use for performance improvement at the time of initial endorsement, provide a credible rationale that describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.**
Not Applicable.

4c. Unintended Consequences

The benefits of the performance measure in facilitating progress toward achieving high-quality, efficient healthcare for individuals or populations outweigh evidence of unintended negative consequences to individuals or populations (if such evidence exists).

4c.1. **Were any unintended negative consequences to individuals or populations identified during testing; OR has**

evidence of unintended negative consequences to individuals or populations been reported since implementation? If so, identify the negative unintended consequences and describe how benefits outweigh them or actions taken to mitigate them.

Not Applicable.

5. Comparison to Related or Competing Measures

If a measure meets the above criteria and there are endorsed or new related measures (either the same measure focus or the same target population) or competing measures (both the same measure focus and the same target population), the measures are compared to address harmonization and/or selection of the best measure.

5. Relation to Other NQF-endorsed Measures
Are there related measures (conceptually, either same measure focus or target population) or competing measures (conceptually both the same measure focus and same target population)? If yes, list the NQF # and title of all related and/or competing measures.
Yes

5.1a. **List of related or competing measures (selected from NQF-endorsed measures)**
0469 : PC-01 Elective Delivery

5.1b. **If related or competing measures are not NQF endorsed please indicate measure title and steward.**

5a. Harmonization

The measure specifications are harmonized with related measures;
OR
The differences in specifications are justified

5a.1. **If this measure conceptually addresses EITHER the same measure focus OR the same target population as NQF-endorsed measure(s):**
Are the measure specifications completely harmonized?
Yes

5a.2. **If the measure specifications are not completely harmonized, identify the differences, rationale, and impact on interpretability and data collection burden.**
The measures are completely harmonized to the extent possible, given the fact that the data source for #0469 is the paper medical record, and the data source for #2829 is the electronic health record.

5b. Competing Measures

The measure is superior to competing measures (e.g., is a more valid or efficient way to measure);
OR
Multiple measures are justified.

5b.1. **If this measure conceptually addresses both the same measure focus and the same target population as NQF-endorsed measure(s):**
Describe why this measure is superior to competing measures (e.g., a more valid or efficient way to measure quality); OR provide a rationale for the additive value of endorsing an additional measure. (Provide analyses when possible.)
Not Applicable.

Appendix

A.1 Supplemental materials may be provided in an appendix. All supplemental materials (such as data collection instrument or methodology reports) should be organized in one file with a table of contents or bookmarks. If material pertains to a specific submission form number, that should be indicated. Requested information should be provided in the submission form and required attachments. There is no guarantee that supplemental materials will be reviewed.

Available at measure-specific web page URL identified in S.1 **Attachment:**

Contact Information

Co.1 **Measure Steward (Intellectual Property Owner):** The Joint Commission
Co.2 **Point of Contact:** Ann, Watt, awatt@jointcommission.org, 630-792-5944-
Co.3 **Measure Developer if different from Measure Steward:** The Joint Commission
Co.4 **Point of Contact:** Michelle, Dardis, mdardis@jointcommission.org, 630-792-5066-

Additional Information

Ad.1 **Workgroup/Expert Panel involved in measure development**
Provide a list of sponsoring organizations and workgroup/panel members' names and organizations. Describe the members' role in measure development.
The technical advisory panel (TAP) members determined priority areas that could be evaluated to improve care related to perinatal care during the development timeframe. After implementation, minor revisions, acknowledged by TAP representatives, were made to improve clarity. Members are:

Michael Ross, MD, MPH (Chair)
Harbor-UCLA Medical Center
Torrance, CA

Wanda Barfield, MD, MPH
Centers for Disease Control and Prevention
Atlanta, GA

Kenneth E. Brown, MD, MBA, FACOG, FACHE
Woman´s Hospital
Lafayette, LA

Martin McCaffrey, MD
UNC North Carolina Children's Hospital
Chapel Hill, NC

Cathy Collins-Fulea, MSN, CNM
Henry Ford Hospital
Detroit, MI

Janet H. Muri, MBA
National Perinatal Information Center/
Quality Analytic Services
Providence, RI

Kathleen Simpson, PhD, RNC, FAAN
St. John's Mercy Medical Center
St. Louis, MO

Michael Socol, MD
Northwestern University Medical School
Chicago, IL

Rebecca Zimmermann, MPP
America's Health Insurance Plans
Washington, DC

Measure Developer/Steward Updates and Ongoing Maintenance
Ad.2 **Year the measure was first released:** 2012
Ad.3 **Month and Year of most recent revision:** 05, 2015
Ad.4 **What is your frequency for review/update of this measure?** Annual
Ad.5 **When is the next scheduled review/update for this measure?** 04, 2016

Ad.6 **Copyright statement:** Measure specifications are in the Public Domain

LOINC(R) is a registered trademark of the Regenstrief Institute.

This material contains SNOMED Clinical Terms (R) (SNOMED CT(c)) copyright 2004-2014 International Health Terminology Standards Development Organization. All rights reserved.

Ad.7 **Disclaimers:** These performance measures are not clinical guidelines and do not establish a standard of medical care, and have not been tested for all potential applications. The measures and specifications are provided without warranty.

Ad.8 **Additional Information/Comments:** Release notes are published each year to reflect changes made in the annual update of the electronic specifications. For changes made in the most recent update, please refer to the "2014 eCQM Technical Release Notes Update June 2015" document, found at:
https://www.cms.gov/Regulations-and-Guidance/Legislation/EHRIncentivePrograms/Downloads/EHandEPTRNs.pdf

MEASURE WORKSHEET

This document summarizes the evaluation of the measure as it progresses through NQF's Consensus Development Process (CDP). The information submitted by measure developers/stewards is included after the Brief Measure Information, Preliminary Analysis, and Pre-meeting Public and Member Comments sections.

To navigate the links in the worksheet: Ctrl + click link to go to the link; ALT + LEFT ARROW to return

Brief Measure Information

NQF #: 480:2830
Measure Title: PC-05 Exclusive Breast Milk Feeding
Measure Steward: The Joint Commission
Brief Description of Measure: PC-05 assesses the number of newborns exclusively fed breast milk during the newborn´s entire hospitalization. This measure is a part of a set of five nationally implemented measures that address perinatal care (PC-01: Elective Delivery, PC-02: Cesarean Section, PC-03: Antenatal Steroids, PC-04: Health Care-Associated Bloodstream Infections in Newborns). PC-05, Exclusive Breast Milk Feeding, is one of two measures in this set that have been reengineered as eCQMs and are included in the EHR Incentive Program and Hospital Inpatient Quality Reporting Program.
Developer Rationale: Increasing the number of newborns who are exclusively fed breast milk for the first six months of life remains a major goal of the WHO, DHHS, AAP and ACOG. Guidelines for the promotion of breast milk feeding are available from the CDC to assist hospitals in establishing successful interventions to improve exclusive breast milk feeding rates in newborns. Breast milk feeding results in numerous health benefits for both mother and newborn. Breastfeeding is associated with decreased risk for many early-life diseases and conditions, including otitis media, respiratory tract infections, atopic dermatitis, gastroenteritis, type 2 diabetes, sudden infant death syndrome, and obesity. Breastfeeding also is associated with health benefits to women, including decreased risk for type 2 diabetes, ovarian cancer, and breast cancer
The measure will assist health care organizations (HCOs) to track evidence of an increase in the number of newborns who were exclusively fed breast milk during the birth hospitalization.

Numerator Statement: Newborns that were fed breast milk only since birth
Denominator Statement: Single term newborns discharged from the hospital who did not have a diagnosis of galactosemia, were not subject to parenteral nutrition, and had a length of stay of less than or equal to 120 days
Denominator Exclusions:
- Newborns who were admitted to the Neonatal Intensive Care Unit (NICU)
- Newborns who were transferred to an acute care facility
- Newborns who expired during the hospitalization

Measure Type: Process
Data Source: Electronic Clinical Data, Electronic Clinical Data : Electronic Health Record
Level of Analysis: Facility, Population : National

IF Endorsement Maintenance – Original Endorsement Date: Most Recent Endorsement Date:

New Measure -- Preliminary Analysis

Criteria 1: Importance to Measure and Report
1a. Evidence .
1a. Evidence. The evidence requirements for a health outcomes measure include providing rationale that supports the

relationship of the health outcome to processes or structures of care. The guidance for evaluating the clinical evidence asks if the relationship between the measured health outcome and at least one clinical action is identified and supported by the stated rationale.

This measure is the new eMeasure version of measure 0480. The information provided for Evidence and Opportunity for Improvement is identical to that submitted for 0480. Measure 0480 will be discussed first – the ratings for evidence and opportunity for improvement will automatically be assigned to this eMeasure without further discussion.

Summary of evidence:

- **Systematic Review of the evidence specific to this measure?** ☒ **Yes** ☐ **No**
- **Quality, Quantity and Consistency of evidence provided?** ☒ **Yes** ☐ **No**
- **Evidence graded?** ☐ **Yes** ☒ **No**

Summary of prior review in 2012

The evidence supporting this measure is based on clinical practice guidelines from the Academy of Breastfeeding Medicine, based on recommendations from the Office on Women´s Health of the U.S. Department of Health and Human Services, the American Academy of Pediatrics, the American College of Obstetricians and Gynecologists, the American Academy of Family Physicians, the World Health Organization (WHO), the Academy of Breastfeeding Medicine, and the UNICEF/WHO evidence-based "Ten Steps to Successful Breastfeeding." The recommendations were based on a review of the literature, which indicates that there are numerous health benefits to breastfeeding for both mother and newborn. The developer states that over 27,000 articles on this topic were published between 1980 and 2012, and the literature includes "900 studies which examine outcomes from breast-feeding with reductions in asthma, diarrheal illness, and childhood obesity being the most important health benefits."

Changes to evidence from last review

☒ **The developer attests that there have been no changes in the evidence since the measure was last evaluated.**

☐ **The developer provided updated evidence for this measure:**

Updates: N/A

Guidance from the Evidence Algorithm

Process measure (Box 1) → Systematic review (Box 3) → QQC present (Box 4) →SR concludes high quality evidence

Preliminary rating for evidence: ☒ **Pass** ☐ **No Pass**
This measure will be assigned the same rating for evidence as measure 0480 without further discussion.

1b. Gap in Care/Opportunity for Improvement and 1b. disparities

1b. Performance Gap. The performance gap requirements include demonstrating quality problems and opportunity for improvement.

No data specific to the eMeasure is submitted.

The developer believes a goal of 70% for exclusive breastmilk feeding is achievable. Rates for exclusive breast milk feeding remain below 50% for over half of the hospitals reporting data. Only the 90th percentile hospitals (74.3%) are performing above the goal of 70%. The developer reports the following data:

Baseline 2Q 2010: 165 hospitals, average measure rate of 40.9% (n=54,630 patients).

Beginning 2014, all hospitals with >1110 births/year were required to report; 1386 hospitals reported the data with an

average rate of 49.4% (n=728,157 patients).

	2011	2012	2013	2014
# hospitals	166	170	197	1386
# newborns	69,613	76,952	91,011	728,157
National aggregate rate	45.9%	51.1%	53.6%	49.4%
10^{th}-25^{th} -90^{th} %tile	19.9% - 34.8% - 62.8%	21.1% - 41.7% - 80.7%	28.4% - 42.3% - 79.3%	22.0% - 35.1% - 74.3%
Mean hospital rate (SD)	0.49335 (0.21348)	0.55872 (0.20359)	0.5632 (0.19707)	0.48724 (0.19475)

Beginning in January 2016 hospitals with >300 births/year are required to report – an additional 821 (approximately 80% of all birthing hospitals).

Disparities

No disparities information from the use of the measure is provided.

The developer provides literature which states that in a study of 307 mothers, "exclusive in-hospital breast milk feeding was reported by 54.2% of white, 38.7% of black, 54.0% of asian, and 44.7% of hispanic (p = 0.063), and among these, only 55.6%, 50.0%, 58.9%, and 19.1%, respectively, maintained exclusive breast milk feeding during the first postpartum month (p < 0.02). The rate of exclusive breast milk feeding at the end of the first month was 10.5%, 15.8%, 20.7%, and 3.9%, respectively, for the white, black, asian, and hispanic mothers whose infants received partial or no breastfeeding in-hospital."

The developer also provided CDC data stating that "from 2000-2004 the rates of exclusive breastfeeding were significantly lower among black infants (compared with white infants) and infants born to unmarried mothers (compared with married mothers). Additionally, older age, urban residence, higher education, and higher income of mothers all were positively associated with exclusive breast milk feeding."

Questions for the Committee:

- *Is there a gap in care that warrants a national performance measure?*
- *If no disparities information is provided, are you aware of evidence that disparities exist in this area of healthcare?*

Preliminary rating for opportunity for improvement: ☒ High ☐ Moderate ☐ Low ☐ Insufficient
This measure will be assigned the same rating for gap as measure 0480 without further discussion.

Committee pre-evaluation comments

Criteria 1: Importance to Measure and Report (including 1a, 1b, 1c)

1a. Evidence to Support Measure Focus

Comments:

Agree with preliminary Pass. Exceeds burden of proof.

1b. Performance Gap

Comments:

Agree with high opportunity for improvement

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability

2a1. Reliability Specifications

2a1. Specifications requires the measure, as specified, to produce consistent (reliable) and credible (valid) results about the quality of care when implemented.

Data source(s): EHR

Specifications: HQMF specifications are provided – see technical review

eMeasure Technical Advisor review:

Submitted measure is an HQMF compliant eMeasure	The submitted eMeasure specifications follow the industry accepted format for eMeasure (HL7 Health Quality Measures Format (HQMF)). **HQMF specifications** ☒ **Yes** ☐ **No**
Documentation of HQMF or QDM limitations	N/A – All components in the measure logic of the submitted eMeasure are represented using the HQMF and QDM
Value Sets	The submitted eMeasure specifications use existing value sets when possible and uses new value sets that have been vetted through the VSAC
Measure logic is unambiguous	Submission includes test results from a simulated data set demonstrating the measure logic can be interpreted precisely and unambiguously Demonstrated through Bonnie testing
Feasibility Testing	The submission contains a feasibility assessment that addresses data element feasibility and follow-up with the measure developer indicates that the measure logic is feasible. This is a legacy eMeasure included in the Meaningful Use program. The developer submitted Bonnie testing results to establish feasibility, and provided an interpretation of the testing process and results in the measure testing attachment

Questions for the Committee:

- *Are the eMeasure specification fully aligned with the chart-based measure 0469?*

2a2. Reliability Testing Testing attachment

2a2. Reliability testing demonstrates if the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise enough to distinguish differences in performance across providers.

SUMMARY OF TESTING

Reliability testing level ☐ **Measure score** ☐ **Data element** ☐ **Both**

Reliability testing performed with the data source and level of analysis indicated for this measure ☐ **Yes** ☐ **No**

Method(s) of reliability testing - Data element validity testing will also count for reliability -see validity section

Results of reliability testing - see validity section

Preliminary rating for reliability: ☐ **High** ☐ **Moderate** ☐ **Low** ☐ **Insufficient**
The rating for reliability will then be the same as validity.

2b. Validity

2b1. Validity: Specifications

2b1. Validity Specifications. This section should determine if the measure specifications are consistent with the evidence.

Specifications consistent with evidence in 1a. ☒ **Yes** ☐ **Somewhat** ☐ **No**
Specification not completely consistent with evidence

Question for the Committee:
- *Are the specifications consistent with the evidence?*

2b2. Validity testing

2b2. Validity Testing should demonstrate the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality.

SUMMARY OF TESTING
Validity testing level ☐ **Measure score** ☒ **Data element testing against a gold standard** ☐ **Both**

Method of validity testing of the measure score:
- ☐ **Face validity only**
- ☐ **Empirical validity testing of the measure score**

NQF supports the development and use of eMeasures. Until more widespread use of eMeasures provides data to formally test eMeasures, NQF will accept testing in a simulated dataset, such as the BONNIE tool, to meet the requirements for measure testing.

Validity testing method:
- BONNIE testing -simulated data set of 28 patients – see attachment
- The developer further describes the BONNIE testing "Cases were used to test the validity of each data element and timing relationship in the measure. Patient characteristics such as age, diagnosis, and length of stay were pre-determined to provide a variety of scenarios that adequately tested for patients passing each data element and failing each data element. Data included in cases and tested for this measure included diagnoses, differing types of delivery and uterine procedures, gestational ages, medications, patient orders, age, length of stay, and clinical observations, such as time of delivery."

Validity testing results:
- The developer reports that "All 28 cases passed or failed as expected based on the data included in the case, confirming the measure logic is accurate and valid."
- "Bonnie testing includes negative and positive testing of each data element in the measure. Positive testing ensures patients expected to be included in the measure are included. Negative testing ensures that patients who do not meet the data criteria are not included in the measure."

Questions for the Committee:
- *Is the test sample adequate to generalize for widespread implementation?*

- *Do the results demonstrate sufficient validity so that conclusions about quality can be made?*
- *Do you agree that the score from this measure as specified is an indicator of quality?*
- *Other specific question of the validity testing?*

2b3-2b7. Threats to Validity

2b3. Exclusions:

- The developer states the "Exclusions in the eMeasure align with the chart-based version of the measure, and are clinically necessary for the interpretation of the measure. No additional exclusion analyses were done on the eMeasure."

Questions for the Committee:

- *Are the exclusions consistent with the evidence?*
- *Is it reasonable to assume that the impact of exclusions will be similar for the eMeasure and the chart-abstracted measure?*

2b4. Risk adjustment: **Risk-adjustment method** ☒ **None** ☐ **Statistical model** ☐ **Stratification**

2b5. Meaningful difference *(can statistically significant and clinically/practically meaningful differences in performance measure scores can be identified)*:

No information provided for the eMeasure. Data was provided for the paper-based measure 0480.

Question for the Committee:

- *Does this measure identify meaningful differences about quality?*

2b6. Comparability of data sources/methods:
N/A

2b7. Missing Data

- The developer reports "Data not present in the structured field from which the measure draws will not be included in the measure calculation. In the Bonnie testing environment, missing data are tested as an expected "Fail" for that data element, and actual performance (whether or not the case fails) is compared to expectations to ensure missing data impacts data element and overall measure calculation as expected."

Preliminary rating for validity: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**
Guidance from Validity Algorithm: Specifications consistent with evidence (Box 1) → threats to validity empirically assessed to some extent (Box 2)→ empirical testing of data elements using BONNIE tool (Box 6) →acceptable BONNIE testing 3))→Moderate

Committee pre-evaluation comments

Criteria 2: Scientific Acceptability of Measure Properties (including all 2a, 2b, and 2d)

2a1. & 2b1. Specifications
Comments:
High reliability demonstrated, capable of larger demonstrations using existent data- however, BONNIE testing med requirements.
BONNIE testing performed.

2a2. Reliability Testing
Comments:
"acceptable."

2b2. Validity Testing
Comments:
"acceptable."

2b3. Exclusions Analysis
2b4. Risk Adjustment/Stratification for Outcome or Resource Use Measures
2b5. Identification of Statistically Significant & Meaningful Differences In Performance
2b6. Comparability of Performance Scores When More Than One Set of Specifications
2b7. Missing Data Analysis and Minimizing Bias
Comments:
Aligned with existent chart based measure.

Criterion 3. Feasibility

3. Feasibility is the extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

- This is an eMeasure for use in EHRs.
- The developer summarizes "Cases were used to test the validity of each data element and timing relationship in the measure. Patient characteristics such as age, diagnosis, and length of stay were pre-determined to provide a variety of scenarios that adequately tested for patients passing each data element and failing each data element. Data included in cases and tested for this measure included diagnoses, differing types of delivery and uterine procedures, gestational ages, medications, patient orders, age, length of stay, and clinical observations, such as time of delivery."
- BONNIE testing supports the feasibility of this measure.

Questions for the Committee:

- *Are the required data elements routinely generated and used during care delivery?*
- *Are the required data elements available in electronic form, e.g., EHR or other electronic sources?*
- *Is the data collection strategy ready to be put into operational use?*
- *If an eMeasure, does the eMeasure Feasibility Score Card demonstrate acceptable feasibility in multiple EHR systems and sites?*

Preliminary rating for feasibility: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 3: Feasibility

3a. Byproduct of Care Processes
3b. Electronic Sources
3c. Data Collection Strategy
Comments:
Yes, required data elements are routinely generated, are electronically available, and should be put to operational use. So much so- I am sure BONNIE testing was needed.

Criterion 4: Usability and Use

4. Usability and Use evaluate the extent to which audiences (e.g., consumers, purchasers, providers, policymakers) use or could use performance results for both accountability and performance improvement activities.

Current uses of the measure
Publicly reported? ☒ **Yes** ☐ **No**

Current use in an accountability program? ☒ **Yes** ☐ **No**
OR
Planned use in an accountability program? ☐ **Yes** ☐ **No**

Accountability program details

- Payment Program --Hospital Inpatient Quality Reporting Program https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/HospitalQualityInits/HospitalRHQDAPU.html
- Regulatory and Accreditation Programs Hospital Accreditation Program http://jointcommission.org
- EHR Incentive Program https://www.cms.gov/Regulations-and-Guidance/Legislation/EHRIncentivePrograms/index.html?redirect=/ehrincentiveprograms

Improvement results Data specific to the eMeasure is not available. See measure 0480.

- The developer states that while the chart-abstracted version of this measure has indicated improvement over time, this measure is a legacy eCQM that is currently in use in the Hospital Inpatient Quality Reporting Program (HIQR) and the EHR Incentive Program. At present, no performance data are currently available.

Unexpected findings (positive or negative) during implementation

- Developer did not identify unexpected findings during implementation.

Potential harms

- Developer did not identify any unintended consequences related to this measure.

Feedback:

***Questions for the Committee**:*

- *How can the performance results be used to further the goal of high-quality, efficient healthcare?*
- *Do the benefits of the measure outweigh any potential unintended consequences?*

Preliminary rating for usability and use: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 4: Usability and Use

4a. Accountability and Transparency
4b. Improvement
4c. Unintended Consequences
Comments:
Yes- HIQR, legacy eCQM.

Criterion 5: Related and Competing Measures

Related or competing measures

This is the eMeasure version of measure 0480.

Harmonization

Pre-meeting public and member comments

- I am writing to express my strong support for continued endorsement of PC-05, Exclusive Breast Milk Feeding. Seminal research from the California Maternity Care Quality Collaborative demonstrated substantial variation in supplementation of breastfed infants among maternity centers. Moreover, national data confirm that there is wide variation in the use of formula among breastfed infants in the first 2 days of life[1], ranging from 6.1% in Montana to 34.9% in New Jersey. This variation among states suggests that overutilization of formula occurs in many maternity hospitals.
  However, just as some infants require delivery via cesarean, some breastfeeding dyads require formula supplementation. It is therefore essential that implementation of PC05 occurs within a context that provides appropriate support for family-centered decision-making and transitions to outpatient support. The AAP recommends that all breastfeeding newborns be seen within 48 to 72h of discharge from the maternity center[2]. Because some families may initiate breastfeeding after leaving the hospital, it may be prudent to schedule all newborns for a 48 to 72h visit to establish care with a pediatric provider. It may be useful to consider a quality measure for the proportion of infants seen by a health professional, either in the office or for a home visit, within 48 to 72h of discharge.
  Of note, the Baby Friendly Hospital Initiative includes a metric for exclusive breast milk feeding as one of its metrics for certification. Differences exist between PC-05 and the BFHI measure, increasing reporting burden for maternity centers. It would be helpful if BFHI and NQF could work together to develop a common metric for measuring exclusive breast milk feeding.
  Evidence continues to accrue that there is no replacement for mother's milk[3]. We can enable families to achieve optimal infant feeding by reducing iatrogenic formula supplementation during the maternity stay, and by ensuring careful follow-up for all families in the early days of life.

  1. Centers for Disease Control and Prevention. 2012: Percent of breastfed infants who were supplemented with infant formula within 2 days of life. 2015 [cited 2016 April 4]; Available from: https://nccd.cdc.gov/NPAO_DTM/IndicatorSummary.aspx?category=8&indicator=41.
  2. American Academy of Pediatrics, Breastfeeding and the use of human milk. Pediatrics, 2012. 129(3): p. e827-41.
  3. Victora, C.G., et al., Breastfeeding in the 21st century: epidemiology, mechanisms, and lifelong effect. The Lancet, 2016. 387(10017): p. 475-490.

- **Developer Response:**
  The Joint Commission thanks you for your support of NQF measure # 2830 (2830) PC-05: Exclusive Breast Milk Feeding. We have reached out to Baby Friendly USA in order to get a better understanding of the differences in their exclusive breast milk feeding measure specifications and our measure specifications with a goal of harmonizing to the extent possible.

1. IMPACT, OPPORTUITY, EVIDENCE - IMPORTANCE TO MEASURE AND REPORT

Importance to Measure and Report is a threshold criterion that must be met in order to recommend a measure for endorsement. All three subcriteria must be met to pass this criterion. See guidance on evidence.

Measures must be judged to be important to measure and report in order to be evaluated against the remaining criteria. (evaluation criteria)

1c.1 Structure-Process-Outcome Relationship *(Briefly state the measure focus, e.g., health outcome, intermediate clinical outcome, process, structure; then identify the appropriate links, e.g., structure-process-health outcome; process- health outcome; intermediate clinical outcome-health outcome)*:
The focus of the measure is to increase the number of newborns who are exclusively fed breast milk during the birth hospitalization >> population determined >> population assessed >> newborns exclusively fed breast milk while in the hospital >> reduced morbidity and mortality of for mother and newborn.

1c.2-3 Type of Evidence *(Check all that apply)*:
Clinical Practice Guideline
Systematic review of body of evidence (other than within guideline development)

1c.4 Directness of Evidence to the Specified Measure *(State the central topic, population, and outcomes addressed in the body of evidence and identify any differences from the measure focus and measure target population)*:
The central topic for the measure is promotion of exclusive breast milk feeding of the newborn during the entire birth hospitalization. The evidence shows numerous health benefits for both mothers and newborns. The target population for the performance measure is consistent with the body of evidence supporting the need for improving exclusive breast milk feeding rates.

1c.5 Quantity of Studies in the Body of Evidence *(Total number of studies, not articles)*: 1c.5. Quantity of Studies in the Body of Evidence (Total number of studies, not articles)
The body of literature examining breast feeding with neonatal outcomes is very large with over 27,000 articles published since 1980. 900 studies examine outcomes from breast-feeding with reductions in asthma, diarrheal illness, and childhood obesity being the most important health benefits. Exclusive breast-feeding in the first weeks was the single most important factor. Over 100 studies have examined initial breast feeding as a quality measure. A separate but related evidence base is the World Health Organization and United Nations Children´s Fund (UNICEF) Baby-Friendly Hospital Initiative that specifies Ten Steps to Successful Breastfeeding which identifies hospital practices that impair exclusive breast-feeding (over 200 separate studies).

1c.6 Quality of Body of Evidence (*Summarize the certainty or confidence in the estimates of benefits and harms to patients across studies in the body of evidence resulting from study factors. Please address: a) study design/flaws; b) directness/indirectness of the evidence to this measure (e.g., interventions, comparisons, outcomes assessed, population included in the evidence); and c) imprecision/wide confidence intervals due to few patients or events):* The quality of evidence supporting the promotion and support of exclusive breast milk feeding is quite high with studies published that have involved mother and newborn couplets. As noted numerous RCTs have been conducted over the past decades demonstrating improved health benefits for both mother and newborn. Some of the improved health benefits for newborns include: otitis media risk reduction by 23% (95% CI 9% to 36%), respiratory tract infections risk reduction by 72% (95% CI 46% to 86%), atopic dermatitis risk reduction by 42% (95% CI 8% to 59%), gastroenteritis risk reduction by 64% (95% CI 26% to 82%), type 2 diabetes risk reduction by 39 percent (95% CI 15% to 56%) , sudden infant death syndrome risk reduction by 36 percent (95% CI 19% to 49%), and obesity risk reduction in two studies by 7- 24% (95% CI 14% to 33% and 95% CI 1% to 12%)
No study design flaws were identified during the literature review.

1c.7 Consistency of Results across Studies *(Summarize the consistency of the magnitude and direction of the effect):* Studies spanning the past five decades have consistently demonstrated the health benefits of breast milk feeding for both mother and newborn. Again, some of the improved health benefits for newborns include: otitis media risk reduction by 23% (95% CI 9% to 36%), respiratory tract infections risk reduction by 72% (95% CI 46% to 86%), atopic dermatitis risk reduction by 42% (95% CI 8% to 59%), gastroenteritis risk reduction by 64% (95% CI 26% to 82%), type 2 diabetes risk reduction by 39 percent (95% CI 15% to 56%) , sudden infant death syndrome risk reduction by 36 percent (95% CI 19% to 49%), and obesity risk reduction in two studies by 7- 24% (95% CI 14% to 33% and 95% CI 1% to 12%)

1c.8 Net Benefit *(Provide estimates of effect for benefit/outcome; identify harms addressed and estimates of effect; and net benefit -*

benefit over harms):
As described before, there are no known harms to patients associated with exclusive breast milk feeding. There are numerous studies documenting health benefits to both newborn and mother; therefore, the benefits of this recommended practice outweigh the harms.

1c.9 **Grading of Strength/Quality of the Body of Evidence**. Has the body of evidence been graded? No

1c.10 **If body of evidence graded, identify the entity that graded the evidence including balance of representation and any disclosures regarding bias:**

1c.11 **System Used for Grading the Body of Evidence:** Other

1c.12 **If other, identify and describe the grading scale with definitions:** Although grading of the evidence was not determined during our systematic review, it was determined that the guideline developers accounted for a balanced representation of information, looked beyond one specialty group or discipline, and provided information that was accessible and met the requirements set out in this measure maintenance form.

1c.13 **Grade Assigned to the Body of Evidence:** Not Applicable

1c.14 **Summary of Controversy/Contradictory Evidence:** There is no documented evidence regarding controversy about the benefits of exclusive breast milk feeding for mother and newborn.

1c.15 **Citations for Evidence other than Guidelines*(Guidelines addressed below)*:**

- American College of Obstetricians and Gynecologists (ACOG). (Feb. 2007). Committee on Obstetric Practice and Committee on Health Care for Underserved Women. Breastfeeding: Maternal and Infant Aspects. ACOG Committee Opinion 361.
- Centers for Disease Control and Prevention (CDC). (2011). Hospital support for breastfeeding: Preventing obesity begins in hospitals. CDC Vital Signs, Retrieved September 26, 2011 at: http://www.cdc.gov/VitalSigns/pdf/2011-08-vitalsigns.pdf
- Ip S, Chung M, Raman G, et al. Breastfeeding and maternal and infant health outcomes in developed countries. Rockville, MD: US Department of Health and Human Services; 2007. Retrieved on September 27, 2011 at: http://www.ahrq.gov/downloads/pub/evidence/pdf/brfout/brfout.pdf.
- Kramer, M.S. & Kakuma, R. (2002).Optimal duration of exclusive breastfeeding. [107 refs] Cochrane Database of Systematic Reviews. (1):CD003517.
- Shealy, K.R., Li, R., Benton-Davis, S., & Grummer-Strawn, L.M. (2005).The CDC guide to breastfeeding interventions. Atlanta, GA: US Department of Health and Human Services, CDC. Available at: http://www.cdc.gov/breastfeeding/pdf/breastfeeding_interventions.pdf
- US Department of Health and Human Services (DHHS). (2000). Healthy People 2010. Washington, DC. Retrieved on Setember 26, 2011 at: http://www.healthypeople.gov/2010
- US Department of Health and Human Services (DHHS). (2010). Healthy People 2020. Washington, DC. Retrieved on September 26, 2011 at: http://www.healthypeople.gov/2020
- World Health Organization (WHO). Indicators for assessing breastfeeding practices. Geneva, Switzerland: World Health Organization; 1991. Retrieved on September 27, 2011 at: http://www.who.int/child-adolescent-health/new_publications/nutrition/who_cdd_ser_91.14.pdf.

1c.16 **Quote verbatim, <u>the specific guideline recommendation</u>** *(Including guideline # and/or page #)*:
The following major recommendations are included in the Academy of Breastfeeding Medicine Protocol # 7on pages 173-177:
Policy Statements
1. The "name of institution" staff will actively support breastfeeding as the preferred method of providing nutrition to infants. A multidisciplinary, culturally appropriate team comprising hospital administrators, physician and nursing staff, lactation consultants and specialists, nutrition staff, other appropriate staff, and parents shall be established and maintained to identify and eliminate institutional barriers to breastfeeding. On a yearly basis, this group will compile and evaluate data relevant to breastfeeding support services and formulate a plan of action to implement needed changes. (III)
2. A written breastfeeding policy will be developed and communicated to all health care staff. The "name of institution" breastfeeding policy will be reviewed and updated biannually using current research as an evidence-based guide. (III)
3. All pregnant women and their support people as appropriate will be provided with information on breastfeeding and counseled on the benefits of breastfeeding, contraindications to breastfeeding, and risk of formula feeding (Academy of Breastfeeding Medicine Protocol Committee, "Clinical protocol #19," 2009). (II-1, II-2, III)
4. The woman´s desire to breastfeed will be documented in her medical record. (III)

5. Mothers will be encouraged to exclusively breastfeed unless medically contraindicated. The method of feeding will be documented in the medical record of every infant. (Exclusive breastfeeding is defined as providing breast milk as the sole source of nutrition.) Exclusively breastfed babies receive no other liquids or solids, with the exception of oral medications prescribed by a medical care provider for the infant.) (II-1, II-2, III)
6. At birth or soon thereafter all newborns, if baby and mother are stable, will be placed skin-to-skin with the mother. Skin-to-skin contact involves placing the naked baby prone on the mother´s bare chest. The infant and mother can then be dried and remain together in this position with warm blankets covering them as appropriate. Mother–infant couples will be given the opportunity to initiate breastfeeding within 1 hour of birth. Post-cesarean-birth babies will be encouraged to breastfeed as soon as possible, potentially in the operating room or recovery area (see Table 1 in the original guideline document). The administration of vitamin K and prophylactic antibiotics to prevent ophthalmia neonatorum should be delayed for the first hour after birth to allow uninterrupted mother–infant contact and breastfeeding (Academy of Breastfeeding Medicine Protocol Committee, "ABM clinical protocol #3," 2009; Mikiel-Kostyra, Mazur, & Boltruszko, 2002; Righard & Alade, 1990). (II-1)
7. Breastfeeding mother–infant couples will be encouraged to remain together throughout their hospital stay, including at night (rooming-in). Skin-to-skin contact will be encouraged as much as possible. (II-1)
8. Breastfeeding assessment, teaching, and documentation will be done on each shift and whenever possible with each staff contact with the mother. Each feeding will be documented, including latch, position, and any problems encountered, in the infant´s medical record. For feedings not directly observed, maternal report may be used. Every shift, a direct observation of the baby´s position and latch-on during feeding will be performed and documented. (II-1, II-2, III)
9. Mothers will be encouraged to utilize available breastfeeding resources including classes, written materials, and video presentations, as appropriate. If clinically indicated, the healthcare professional or nurse will make a referral to a lactation consultant or specialist for additional education or assistance. (II-1, II-2, III)
10. Breastfeeding mothers will be instructed about:
a. Proper positioning and latch on
b. Nutritive suckling and swallowing
c. Milk production and release
d. Frequency of feeding/feeding cues
e. Hand expression of breast milk and use of a pump if indicated
f. How to assess if infant is adequately nourished
g. Reasons for contacting the healthcare professional

These skills will be taught to primiparous and multiparous women, provided in written form (Eidelman, Hoffmann, & Kaitz, 1993), and reviewed before the mother goes home. (II-1, II-2, III)
11. Parents will be taught that breastfeeding infants, including cesarean-birth babies, should be put to breast at least 8 to 12 times each 24 hours, with some infants needing to be fed more frequently. Infant feeding cues (e.g., increased alertness or activity, mouthing, or rooting) will be used as indicators of the baby´s readiness for feeding. Breastfeeding babies will be breastfed at night. (II-1, II-2, III)
12. Time limits for breastfeeding on each side will be avoided. Infants can be offered both breasts at each feeding but may be interested in feeding only on one side at a feeding during the early days. (II-1, II-2, III)
13. No supplemental water, glucose water, or formula will be given unless specifically ordered by a healthcare professional (e.g., physician, certified nurse midwife, or nurse practitioner) or by the mother´s documented and informed request. Prior to non-medically indicated supplementation, mothers will be informed of the risks of supplementing. The supplement should be fed to the baby by cup if possible and will be no more than 10 to 15 mL (per feeding) in a term baby (during the first 1 to 2 days of life). Alternative feeding methods such as syringe or spoon feeding may also be used; however, these methods have not been shown to be effective in preserving breastfeeding. Bottles will not be placed in a breastfeeding infant´s bassinet (Howard et al., 2003; Howard et al., 1999; Marinelli, Burke, & Dodd, 2001). (II-1, II-2)
14. This institution does not give group instruction in the use of formula. Those parents who, after appropriate counseling, choose to formula feed their infants will be provided individual instruction.
15. Pacifiers will not be given to normal full-term breastfeeding infants. The pacifier guidelines at "name of institution" state that preterm infants in the Neonatal Intensive Care or Special Care Unit or infants with specific medical conditions (e.g., neonatal abstinence syndrome) may be given pacifiers for non-nutritive sucking. Newborns undergoing painful procedures (e.g., circumcision) may be given a pacifier as a method of pain management during the procedure. The infant will not return to the mother with the pacifier. "Name of institution" encourages "pain-free newborn care," which may include breastfeeding during the heel stick procedure for the newborn metabolic screening tests (Gray et al., 2002). (I)
16. Routine blood glucose monitoring of full-term healthy appropriate-for-gestational age infants is not indicated. Assessment for clinical signs of hypoglycemia and dehydration will be ongoing (Wight, Marinelli, & Academy of Breastfeeding Medicine Clinical Protocol Committee, 2006). (I)
17. Antilactation drugs will not be given to any postpartum mother. (I)

18. Routine use of nipple creams, ointments, or other topical preparations will be avoided unless such therapy has been indicated for a dermatologic problem. Mothers with sore nipples will be observed for latch-on techniques and will be instructed to apply expressed colostrum or breast milk to the areola/nipple after each feeding. (III)
19. Nipple shields or bottle nipples will not be routinely used to cover a mother´s nipples, to treat latch-on problems, or to prevent or manage sore or cracked nipples or used when a mother has flat or inverted nipples. Nipple shields will be used only in conjunction with a lactation consultation and after other attempts to correct the difficulty have failed. (III)
20. After 24 hours of life, if the infant has not latched on or fed effectively, the mother will be instructed to begin to massage her breasts and hand express colostrum into the baby´s mouth during feeding attempts. Skin-to-skin contact will be encouraged. Parents will be instructed to watch closely for feeding cues and whenever these are observed to awaken and feed the infant. If the baby continues to feed poorly, hand expression by the mother or a double set-up electric breast pump will be initiated and maintained approximately every 3 hours or a minimum of eight times per day. Any expressed colostrum or mother´s milk will be fed to the baby by an alternative method. The mother will be reminded that she may not obtain much milk or even any milk the first few times she expresses her breasts. Until the mother´s milk is available, a collaborative decision should be made among the mother, nurse, and healthcare professional (e.g., physician/nurse practitioner/certified nurse midwife) regarding the need to supplement the baby. Each day the responsible healthcare professional will be consulted regarding the volume and type of the supplement. Pacifiers will be avoided. In cases of problem feeding, the lactation consultant or specialist will be consulted (Academy of Breastfeeding Medicine Protocol Committee, "ABM clinical protocol #3," 2009). (I, III)
21. If the baby is still not latching on well or feeding well when discharged to home, the feeding/expression/supplementing plan will be reviewed in addition to routine breastfeeding instructions. A follow-up visit or contact will be scheduled within 24 hours. Depending on the clinical situation it may be appropriate to delay discharge of the couplet to provide further breastfeeding intervention, support, and education. (III)
22. All babies should be seen for follow-up within the first few days postpartum. This visit should be with a physician (pediatrician or family physician) or other qualified health care practitioner for a formal evaluation of breastfeeding performance, a weight check, assessment of jaundice and age appropriate elimination: (a) for infants discharged at less than 2 days of age (<48 hours), follow-up at 2 to 4 days of age; (b) for infants discharged between 48 and 72 hours, follow-up at 4 to 5 days of age. Infants discharged after 5 to 6 days may be seen 1 week later.
23. Mothers who are separated from their sick or premature infants will be
a. Instructed on how to use skilled hand expression or the double set up electric breast pump. Instructions will include expression at least eight times per day or approximately every 3 hours for 15 minutes (or until milk flow stops, whichever is greater) around the clock and the importance of not missing an expression session during the night (III)
b. Encouraged to breastfeed on demand as soon as the infant´s condition permits (III)
c. Taught proper storage and labeling of human milk (III)
d. Assisted in learning skilled hand expression or obtaining a double set-up electric breast pump prior to going home (III)
24. Before leaving the hospital (Academy of Breastfeeding Medicine Clinical Protocol Committee, 2007), breastfeeding mothers should be able to:
a. Position the baby correctly at the breast with no pain during the feeding
b. Latch the baby to breast properly
c. State when the baby is swallowing milk
d. State that the baby should be nursed a minimum of eight to 12 times a day until satiety, with some infants needing to be fed more frequently
e. State age-appropriate elimination patterns (at least six urinations per day and three to four stools per day by the fourth day of life)
f. List indications for calling a healthcare professional
g. Manually express milk from their breasts (III)
25. Prior to going home, mothers will be given the names and telephone numbers of community resources to contact for help with breastfeeding, including (the support group or resource recommended by "name of institution").
26. "Name of institution" does not accept free formula or free breast milk substitutes. Nursery or Neonatal Intensive Care Unit discharge bags offered to all mothers will not contain infant formula, coupons for formula, logos of formula companies, or literature with formula company logos.
27. "Name of institution" health professionals will attend educational sessions on lactation management and breastfeeding promotion to ensure that correct, current, and consistent information is provided to all mothers wishing to breastfeed (American Academy of Pediatrics, American Academy of Obstetricians and Gynecologists, 2006).
Contraindications:
Breastfeeding is contraindicated in the following situations:
• Mothers who are human immunodeficiency virus (HIV)-positive in locations where artificial feeding is acceptable, feasible,

affordable, sustainable, and safe (I)
- Mothers currently using illicit drugs (e.g., cocaine, heroin) unless specifically approved by the infant´s healthcare provider on a case-by-case basis (I)
- Mothers taking certain medications. Most prescribed and over-the-counter drugs are safe for the breastfeeding infant. Some medications may make it necessary to interrupt breastfeeding, such as radioactive isotopes, antimetabolites, cancer chemotherapy, some psychotropic medications and a small number of other medications. (III)
- Mothers with active, untreated tuberculosis. A mother can express her milk until she is no longer contagious. (I)
- Infants with galactosemia (I)
- Mothers with active herpetic lesions on her breast(s). Breastfeeding can be recommended on the unaffected breast. (The Infectious Disease Service will be consulted for problematic infectious disease issues.) (I)
- Mothers with onset of varicella within 5 days before or up to 48 hours after delivery, until they are no longer infectious (I)
- Mothers with human T-cell lymphotropic virus type I or type II (I)

1c.17 Clinical Practice Guideline Citation: • Philipp BL, Academy of Breastfeeding Medicine Protocol Committee. ABM clinical protocol #7: model breastfeeding policy (revision 2010). Breastfeed Med 2010 Aug;5(4):173-7.

1c.18 National Guideline Clearinghouse or other URL:
http://www.guideline.gov/content.aspx?id=24013&search=breastfeeding+policy

1c.19 Grading of Strength of Guideline Recommendation. Has the recommendation been graded? Yes

1c.20 If guideline recommendation graded, identify the entity that graded the evidence including balance of representation and any disclosures regarding bias: Academy of Breastfeeding Medicine Protocol Committee

1c.21 System Used for Grading the Strength of Guideline Recommendation: Other

1c.22 If other, identify and describe the grading scale with definitions: The system for categorizing recommendations in this guideline is as follows:

Levels of Evidence
I Evidence obtained from at least one properly randomized controlled trial
II-1 Evidence obtained from well-designed controlled trials without randomization
II-2 Evidence obtained from well-designed cohort or case-control analytic studies, preferably from more than one center or research group
II-3 Evidence obtained from multiple time series with or without the intervention. Dramatic results in uncontrolled experiments (such as the results of the introduction of penicillin treatment in the 1940s) could also be regarded as this type of evidence.
III Opinions of respected authorities, based on clinical experience, descriptive studies and case reports; or reports of expert committees

1c.23 Grade Assigned to the Recommendation: Grading varies from I to III

1c.24 Rationale for Using this Guideline Over Others: This policy is based on recommendations from the most recent breastfeeding policy statements published by the Office on Women´s Health of the U.S. Department of Health and Human Services, the American Academy of Pediatrics, the American College of Obstetricians and Gynecologists, the American Academy of Family Physicians, the World Health Organization (WHO), the Academy of Breastfeeding Medicine, and the UNICEF/WHO evidence-based "Ten Steps to Successful Breastfeeding."
The recommendations were based primarily on a comprehensive review of the existing literature. In cases where the literature does not appear conclusive, recommendations were based on the consensus opinion of the group of experts.

Based on the NQF descriptions for rating the evidence, what was the <u>developer's assessment</u> of the quantity, quality, and consistency of the body of evidence?
1c.25 Quantity: High **1c.26** Quality: High**1c.27** Consistency**:** High

1. Evidence, Performance Gap, Priority – Importance to Measure and Report

Extent to which the specific measure focus is evidence-based, important to making significant gains in healthcare quality, and improving health outcomes for a specific high-priority (high-impact) aspect of healthcare where there is variation in or overall less-than-optimal performance. ***Measures must be judged to meet all subcriteria to pass this criterion and be evaluated against the remaining criteria.***

1a. Evidence to Support the Measure Focus – See attached Evidence Submission Form
0480_Evidence_MSF5.0_Data-635787043120686011-635827511031827598.doc

1b. Performance Gap

Demonstration of quality problems and opportunity for improvement, i.e., data demonstrating:

- considerable variation, or overall less-than-optimal performance, in the quality of care across providers; and/or
- disparities in care across population groups.

1b.1. **Briefly explain the rationale for this measure** *(e.g., the benefits or improvements in quality envisioned by use of this measure)*
Increasing the number of newborns who are exclusively fed breast milk for the first six months of life remains a major goal of the WHO, DHHS, AAP and ACOG. Guidelines for the promotion of breast milk feeding are available from the CDC to assist hospitals in establishing successful interventions to improve exclusive breast milk feeding rates in newborns. Breast milk feeding results in numerous health benefits for both mother and newborn. Breastfeeding is associated with decreased risk for many early-life diseases and conditions, including otitis media, respiratory tract infections, atopic dermatitis, gastroenteritis, type 2 diabetes, sudden infant death syndrome, and obesity. Breastfeeding also is associated with health benefits to women, including decreased risk for type 2 diabetes, ovarian cancer, and breast cancer
The measure will assist health care organizations (HCOs) to track evidence of an increase in the number of newborns who were exclusively fed breast milk during the birth hospitalization.

1b.2. **Provide performance scores on the measure as specified (current and over time) at the specified level of analysis.** *(This is required for endorsement maintenance. Include mean, std dev, min, max, interquartile range, scores by decile. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included). This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*
This measure is a legacy eCQM that is currently included in the Hospital Inpatient Quality Reporting Program (HIQR) and the EHR Incentive Program. At present, no performance data for the electronic version of the measure are yet available.

In CY2016, CMS is requiring organizations participating in HIQR to electronically submit 1 quarter of data on 4 of 28 available eCQMs. This measure is one of the 28.

For more information, refer to CY2016 IPPS Final Rule, located here: https://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/AcuteInpatientPPS/FY2016-IPPS-Final-Rule-Home-Page.html

1b.3. **If no or limited performance data on the measure as specified is reported in 1b2, then provide a summary of data from the literature that indicates opportunity for improvement or overall less than optimal performance on the specific focus of measurement.**
Rates for exclusive breast milk feeding remain below 50% for over half of the hospitals reporting data. A goal of 70% should be achievable based on the Joint Commission's analysis of data for the chart-abstracted version of this measure. The chart-abstracted Perinatal Care (PC) core measures were added as a new core measure set in 2010 for hospitals to select in order to meet their ORYX performance measurement requirement for Joint Commission accreditation purposes at that time, approximately 165 hospitals reported the data with an average measure rate of 40.9% (n=54,630 patients). In January 2014, The Joint Commission required mandatory reporting of the PC measure set for all accredited hospitals with 1100 births or more annually. 1386 hospitals reported the data with an average rate of 49.4% (n=728,157 patients). It is important to note that a performance gap of 48% exists for the 10th percentile of hospitals performing at 22.0% (if 70% is considered goal performance) and that the aggregate rate dropped by 4.2% from the 2013 rate of 53.6%. Only the 90th percentile hospitals (74.3%) are performing above the goal of 70%. The threshold for mandatory reporting was recently lowered to 300 births annually effective January 2016. The new reporting requirement will now capture approximately 80% of all accredited birthing hospitals. As a result, the rates may decrease with the addition of approximately 821 more hospitals reporting data. Below is the specified level of analysis for PC-01 beginning with discharges April 1, 2010 through December 31, 2014.

2Q 2010: 54,630 denominator cases; 22,346 numerator cases; 165 hospitals; 40.9% national aggregate rate; 0.46132 mean of

hospital rates; 0.22274 standard deviation; 75.3% 90th percentile rate; 62.2% 75th percentile rate/upper quartile; 45.9% 50th percentile rate/median rate; 29.4% 25th percentile rate/lower quartile; and 15.5% 10th percentile rate.

CY 2011: 69,613 denominator cases; 31,999 numerator cases; 166 hospitals; 45.9% national aggregate rate; 0.49335 mean of hospital rates; 0.21348 standard deviation; 76.2% 90th percentile rate; 62.8% 75th percentile rate/upper quartile; 50.6% 50th percentile rate/median rate; 34.8% 25th percentile rate/lower quartile; and 19.9% 10th percentile rate.

CY 2012: 76,952 denominator cases; 39,337 numerator cases; 170 hospitals; 51.1% national aggregate rate; 0.55872 mean of hospital rates; 0.20359 standard deviation; 80.7% 90th percentile rate; 72.2% 75th percentile rate/upper quartile; 56.7% 50th percentile rate/median rate; 41.7% 25th percentile rate/lower quartile; and 27.1% 10th percentile rate.

CY 2013: 91,011 denominator cases; 48,758 numerator cases; 197 hospitals; 53.6% national aggregate rate; 0.5632 mean of hospital rates; 0.19707 standard deviation; 79.3% 90th percentile rate; 72.1%% 75th percentile rate/upper quartile; 57.3% 50th percentile rate/median rate; 42.3% 25th percentile rate/lower quartile; and 28.4% 10th percentile rate.

CY 2014: 728,157 denominator cases; 359,633 numerator cases; 1386 hospitals; 49.4% national aggregate rate; 0.48724 mean of hospital rates; 0.19475 standard deviation; 74.3% 90th percentile rate; 62.4% 75th percentile rate/upper quartile; 49.1% 50th percentile rate/median rate; 35.1% 25th percentile rate/lower quartile; and 22.0% 10th percentile rate.

1b.4. **Provide disparities data from the measure as specified (current and over time) by population group, e.g., by race/ethnicity, gender, age, insurance status, socioeconomic status, and/or disability.** *(This is required for endorsement maintenance. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities include.) This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*
A study was conducted by Petrova et al. (2007) to identify the association between the in-hospital feeding pattern and the infant´s post discharge feeding modality during the first month of life in a culturally diverse population of women. Demographic, clinical, and feeding practice data were collected from the medical charts and interviews of mothers conducted in the first month after singleton delivery of healthy term newborns. Among the 307 mothers who completed the study, exclusive in-hospital breast milk feeding was reported by 54.2% of white, 38.7% of black, 54.0% of asian, and 44.7% of hispanic ($p = 0.063$), and among these, only 55.6%, 50.0%, 58.9%, and 19.1%, respectively, maintained exclusive breast milk feeding during the first postpartum month ($p < 0.02$). The rate of exclusive breast milk feeding at the end of the first month was 10.5%, 15.8%, 20.7%, and 3.9%, respectively, for the white, black, asian, and hispanic mothers whose infants received partial or no breastfeeding in-hospital.

Overall, the logistic regression analysis showed significant association between initiation of exclusive breast milk feeding in-hospital and exclusive breast milk feeding at the end of the first month (odds ratio 7.2 and 95% confidence interval 4.0, 12.6). In conclusion, it showed a larger decline in the continuation of exclusive breast milk feeding and the lowest rate of exclusive breast milk feeding at 1 month in the hispanic mothers. Irrespective of race/ethnicity, mothers who practice exclusive breast milk feeding in-hospital are more likely to exclusively fed breast milk throughout the neonatal period.

According to the CDC, from 2000-2004 the rates of exclusive breastfeeding were significantly lower among black infants (compared with white infants) and infants born to unmarried mothers (compared with married mothers). Additionally, older age, urban residence, higher education, and higher income of mothers all were positively associated with exclusive breast milk feeding (CDC, 2007). Hawkins et al. (2015) noted continued disparities among mothers with lower education based on Pregnancy Risk Assessment Monitoring System (PRAMS) data collected from 1999 to 2009.

1b.5. **If no or limited data on disparities from the measure as specified is reported in 1b4, then provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement. Include citations.**

• Centers for Disease Control and Prevention (CDC). (2007). Breastfeeding trends and updated national health objectives for exclusive breastfeeding-United States, birth years 2000-2004. [Journal Article] MMWR - Morbidity & Mortality Weekly Report. 56(30):760-3.
• Hawkins, S., Stern, A., Baum, C. & Gillman, M. (2015). Evaluating the impact of the baby-friendly hospital initiative on breast-feeding rates: a multi-state analysis. Public Health Nutr.18(2):189-97.
• Petrova, A., Hegyi, T., Mehta, R. (2007). Maternal race/ethnicity and one-month exclusive breastfeeding in association with the in-hospital feeding modality. Breastfeeding Medicine: The Official Journal of the Academy of Breastfeeding Medicine. 2(2):92-8.

1c. High Priority (previously referred to as High Impact)
The measure addresses:
- a specific national health goal/priority identified by DHHS or the National Priorities Partnership convened by NQF;

OR

- a demonstrated high-priority (high-impact) aspect of healthcare (e.g., affects large numbers of patients and/or has a substantial impact for a smaller population; leading cause of morbidity/mortality; high resource use (current and/or future); severity of illness; and severity of patient/societal consequences of poor quality).

1c.1. Demonstrated high priority aspect of healthcare
Affects large numbers, Patient/societal consequences of poor quality
1c.2. If Other:

1c.3. Provide epidemiologic or resource use data that demonstrates the measure addresses a high priority aspect of healthcare. List citations in 1c.4.
Exclusive breast milk feeding for the first 6 months of neonatal life has long been the expressed goal of World Health Organization (WHO), US Department of Health and Human Services (DHHS), American Academy of Pediatrics (AAP) and American College of Obstetricians and Gynecologists (ACOG). ACOG reiterated its position in a Committee Opinion on breastfeeding (ACOG, 2007). Additionally, a Cochrane review of two randomized control trials and 18 other studies substantiates the benefits of exclusive breast milk feeding for the first six months of life (Kramer et al., 2002).

Breastfeeding is associated with decreased risk for many early-life diseases and conditions, including otitis media, respiratory tract infections, atopic dermatitis, gastroenteritis, type 2 diabetes, sudden infant death syndrome, and obesity. Breastfeeding also is associated with health benefits to women, including decreased risk for type 2 diabetes, ovarian cancer, and breast cancer (Ip et al., 2007). Exclusive breastfeeding is defined as a newborn receiving only breast milk and no other liquids or solids except for drops or syrups consisting of vitamins, minerals, or medicines (WHO, 1991).

In 2007, Healthy People 2010 objectives for breastfeeding initiation and duration included two new objectives on exclusive breastfeeding to increase the proportion of mothers who exclusively breastfeed their infants through age 3 months to 60% and through age 6 months to 25% [objectives 16-19d and 16-19e] (DHHS, 2000). The Healthy People 2020 objectives for exclusive breastfeeding were continued through age 3 months with a goal of 46.2% and age 6 months with a goal of 25.5% [objectives MICH-21.4 and MICH-21.5]. Also included is the related objective MICH-24: increase the proportion of live births that occur in facilities that provide recommended care for lactating mothers and their babies (DHHS, 2010).

The Centers for Disease Control and Prevention (CDC) developed a Guide to Breastfeeding Interventions in 2005 for the promotion and support of breastfeeding based on detailed input from the spectrum of breastfeeding experts which can be used to help hospitals achieve the Healthy People 2020 objective MICH-24. Institutional changes i.e., attaining Baby Friendly Hospital Initiative status, individual interventions including increased rooming-in of mothers and newborns, early skin to skin contact and discontinuing policies that are not evidence based have been shown to increase breastfeeding initiation and duration rates as well (Shealy et al., 2007). According to the CDC (2011), mothers who want to breastfeed who do not receive hospital support will stop early. The CDC encourages hospitals to partner with Baby-Friendly hospitals to learn how to improve maternity care, use the CDC's Maternity Practice in Infant Nutrition and Care (mPINC) survey data to prioritize changes to improve maternity care practices and stop distributing formula samples and give-aways to breastfeeding mothers.

1c.4. Citations for data demonstrating high priority provided in 1a.3

• American College of Obstetricians and Gynecologists (ACOG). (Feb. 2007). Committee on Obstetric Practice and Committee on Health Care for Underserved Women. Breastfeeding: Maternal and Infant Aspects. ACOG Committee Opinion 361.
• Brown, P., Kaiser, K. & Nailon,R. (2014). Integrating quality improvement and translational research models to increase exclusive breastfeeding. Journal of Obstetric, Gynecologic, and Neonatal Nursing : JOGNN / NAACOG. 43:5,545-553.
• Centers for Disease Control and Prevention (CDC). (2011). Hospital support for breastfeeding: Preventing obesity begins in hospitals. CDC Vital Signs, Retrieved September 26, 2011 at: http://www.cdc.gov/VitalSigns/pdf/2011-08-vitalsigns.pdf
• Chantry, C., Eglash, A. & Labbok, M.(2015). ABM position on breastfeeding. Breastfeeding Medicine. 10(9): 407-411.
• Chantry, C., Dewey,K., Peerson, J., Wagner, E. & Nommsen-Rivers,L. (2014). In-hospital formula use increases early breastfeeding cessation among first-time mothers intending to exclusively breastfeed.The Journal of Pediatrics. 164:6, 1339-45.e5.
• Dias de Oliveira, L., Justo Giugliani,E., Cordova do Espirito Santo,L. & Meirelles Nunes,L. (2014). Counselling sessions increased duration of exclusive breastfeeding: a randomized clinical trial with adolescent mothers and grandmothers. Nutrition Journal. 13, 73-2891-13-73.
• Ip S, Chung M, Raman G, et al. Breastfeeding and maternal and infant health outcomes in developed countries. Rockville, MD: US Department of Health and Human Services; 2007. Retrieved on September 27, 2011 at: http://archive.ahrq.gov/downloads/pub/evidence/pdf/brfout/brfout.pdf

• Kramer, M.S. & Kakuma, R. (2002).Optimal duration of exclusive breastfeeding. [107 refs] Cochrane Database of Systematic Reviews. (1):CD003517.
• Linares, A., Rayens, M., Gomez, M, Gokun,Y. & Dignan,M. (2014). Intention to breastfeed as a predictor of initiation of exclusive breastfeeding in hispanic women. Journal of Immigrant and Minority Health / Center for Minority Public Health. DOI: 10.1007/s10903-014-0049-0.
• Mueffelmann, R., Racine, E., Warren-Findlow,J. & Coffman,M. (2014). Perceived infant feeding preferences of significant family members and mother´s intention to exclusively breastfeed. Journal of Human Lactation : Official Journal of International Lactation Consultant Association, doi: 10.1177/0890334414553941.
• Perrine C.,et al., (2015). How do hospitals measure up in supporting breast-feeding? (2015). MMWR Morb Mortal Wkly Rep. 64:1112
• Ramakrishnan, R.,Oberg, C. & Kirby,R. (2014).The association between maternal perception of obstetric and pediatric care providers´ attitudes and exclusive breastfeeding outcomes. Journal of Human Lactation : Official Journal of International Lactation Consultant Association.3:1,80-87.
• Shealy, K.R., Li, R., Benton-Davis, S., & Grummer-Strawn, L.M. (2005).The CDC guide to breastfeeding interventions. Atlanta, GA: US Department of Health and Human Services, CDC. Available at: http://www.cdc.gov/breastfeeding/pdf/breastfeeding_interventions.pdf
• US Department of Health and Human Services (DHHS). (2000). Healthy People 2010. Washington, DC. Retrieved on September 26, 2011 at: http://www.healthypeople.gov/2010
• US Department of Health and Human Services (DHHS). (2010). Healthy People 2020. Washington, DC. Retrieved on September 26, 2011 at: http://www.healthypeople.gov/2020
• World Health Organization (WHO). Indicators for assessing breastfeeding practices. Geneva, Switzerland: World Health Organization; 1991. Retrieved on September 27, 2011 at: http://www.who.int/child-adolescent-health/new_publications/nutrition/who_cdd_ser_91.14.pdf

1c.5. **If a PRO-PM (e.g. HRQoL/functional status, symptom/burden, experience with care, health-related behaviors), provide evidence that the target population values the measured PRO and finds it meaningful.** *(Describe how and from whom their input was obtained.)*
Not Applicable

2. Reliability and Validity—Scientific Acceptability of Measure Properties

Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. ***Measures must be judged to meet the subcriteria for both reliability and validity to pass this criterion and be evaluated against the remaining criteria.***

2a.1. Specifications The measure is well defined and precisely specified so it can be implemented consistently within and across organizations and allows for comparability. eMeasures should be specified in the Health Quality Measures Format (HQMF) and the Quality Data Model (QDM).

De.5. **Subject/Topic Area** *(check all the areas that apply):*
Perinatal and Reproductive Health, Perinatal and Reproductive Health : Newborn

De.6. **Cross Cutting Areas** *(check all the areas that apply):*
Patient and Family Engagement

S.1. **Measure-specific Web Page** *(Provide a URL link to a web page specific for this measure that contains current detailed specifications including code lists, risk model details, and supplemental materials. Do not enter a URL linking to a home page or to general information.)*
https://www.cms.gov/regulations-and-guidance/legislation/ehrincentiveprograms/ecqm_library.html

S.2a. **If this is an eMeasure**, HQMF specifications must be attached. Attach the zipped output from the eMeasure authoring tool (MAT) - if the MAT was not used, contact staff. (Use the specification fields in this online form for the plain-language description of the specifications)
This is an eMeasure **Attachment:** EH_CMS9v4_NQF0480_BF_ExclusiveBreastFeed.zip

S.2b. **Data Dictionary, Code Table, or Value Sets** *(and risk model codes and coefficients when applicable) must be attached. (Excel or csv file in the suggested format preferred - if not, contact staff)*
Attachment **Attachment:** ExclusiveBreastMilkFeeding_v4_Fri_Nov_13_10.29.14_CST_2015.xls

S.3. **For endorsement maintenance**, please briefly describe any changes to the measure specifications since last endorsement date and explain the reasons.
Changes have been made to the eCQM specifications in order to reflect the revisions to the chart-abstracted measure from which this measure is derived.

S.4. **Numerator Statement** *(Brief, narrative description of the measure focus or what is being measured about the target population, i.e., cases from the target population with the target process, condition, event, or outcome)*
IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
Newborns that were fed breast milk only since birth

S.5. **Time Period for Data** *(What is the time period in which data will be aggregated for the measure, e.g., 12 mo, 3 years, look back to August for flu vaccination? Note if there are different time periods for the numerator and denominator.)*
Episode of care

S.6. **Numerator Details** *(All information required to identify and calculate the cases from the target population with the target process, condition, event, or outcome such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
IF an OUTCOME MEASURE, describe how the observed outcome is identified/counted. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
- Administration of breast milk is represented with the QDM datatype and value set of Substance, Administered: Breast Milk (OID: 2.16.840.1.113883.3.117.1.7.1.30)
- Administration of other dietary intake is represented with Substance, Administered: Dietary Intake Other than Breast Milk (OID: 2.16.840.1.113883.3.117.1.7.1.27)

To access the value sets for the measure, please visit the Value Set Authority Center (VSAC), sponsored by the National Library of Medicine, at this link: https://vsac.nlm.nih.gov/

S.7. **Denominator Statement** *(Brief, narrative description of the target population being measured)*
Single term newborns discharged from the hospital who did not have a diagnosis of galactosemia, were not subject to parenteral nutrition, and had a length of stay of less than or equal to 120 days

S.8. **Target Population Category** *(Check all the populations for which the measure is specified and tested if any):*
Maternal Health

S.9. **Denominator Details** *(All information required to identify and calculate the target population/denominator such as definitions, specific data collection items/responses , code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
Inpatient Encounters are represented using the QDM datatype and value set of Encounter, Performed: Encounter Inpatient (OID: 2.16.840.1.113883.3.117.1.7.1.25). Length of stay is calculated within the measure based on encounter start and end dates.

Single term newborns are represented by the following QDM datatypes and value sets:
o Physical Exam, Performed: Estimated Gestational Age at Birth (Result>=37 weeks) using Estimated Gestational Age at Birth SNOMEDCT Value Set (OID: 2.16.840.1.113762.1.4.1045.47)
o Diagnosis, Active: Single Live Birth using Single Live Birth SNOMEDCT Value Set (2.16.840.1.113883.3.117.1.7.1.25)
o Diagnosis, Active Single Live Born Newborn Born in Hospital using Single Live Born Newborn Born in Hospital Grouping Value Set (2.16.840.1.113883.3.117.1.7.1.26)

- Galactosemia is represented using the QDM datatype and value set of Diagnosis, Active: Galactosemia (OID: 2.16.840.1.113883.3.117.1.7.1.35)
- Parenteral Nutrition is represented using the QDM datatype and value set of Procedure, Performed: Parenteral Nutrition (OID: 2.16.840.1.113883.3.117.1.7.1.38)

S.10. **Denominator Exclusions** *(Brief narrative description of exclusions from the target population)*
- Newborns who were admitted to the Neonatal Intensive Care Unit (NICU)
- Newborns who were transferred to an acute care facility
- Newborns who expired during the hospitalization

S.11. **Denominator Exclusion Details** *(All information required to identify and calculate exclusions from the denominator such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
NICU admissions, transfers to another facility, and patient expiration are all represented in QDM as attributes of the inpatient encounter.
o facility location: Neonatal Intensive Care Unit (NICU) (OID: 2.16.840.1.113883.3.117.1.7.1.75)
o discharge status: Patient Expired (OID: 2.16.840.1.113883.3.117.1.7.1.309)
o discharge status: Discharge to Acute Care Facility (OID: 2.16.840.1.113883.3.117.1.7.1.87)

S.12. **Stratification Details/Variables** *(All information required to stratify the measure results including the stratification variables, definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format with at S.2b)*
Not Applicable

S.13. **Risk Adjustment Type** (Select type. Provide specifications for risk stratification in S.12 and for statistical model in S.14-15)
No risk adjustment or risk stratification
If other:

S.14. **Identify the statistical risk model method and variables** *(Name the statistical method - e.g., logistic regression and list all the risk factor variables. Note - risk model development and testing should be addressed with measure testing under Scientific Acceptability)*
Not Applicable

S.15. **Detailed risk model specifications** *(must be in attached data dictionary/code list Excel or csv file. Also indicate if available at measure-specific URL identified in S.1.)*
Note: Risk model details (including coefficients, equations, codes with descriptors, definitions), should be provided on a separate

worksheet in the suggested format in the Excel or csv file with data dictionary/code lists at S.2b.

S.15a. **Detailed risk model specifications** (*if not provided in excel or csv file at S.2b*)

S.16. **Type of score:**
Rate/proportion
If other:

S.17. **Interpretation of Score** *(Classifies interpretation of score according to whether better quality is associated with a higher score, a lower score, a score falling within a defined interval, or a passing score)*
Better quality = Higher score

S.18. **Calculation Algorithm/Measure Logic** *(Describe the calculation of the measure score as an ordered sequence of steps including identifying the target population; exclusions; cases meeting the target process, condition, event, or outcome; aggregating data; risk adjustment; etc.)*
See attached HQMF file

S.19. **Calculation Algorithm/Measure Logic Diagram URL or Attachment** *(You also may provide a diagram of the Calculation Algorithm/Measure Logic described above at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
Available at measure-specific web page URL identified in S.1

S.20. **Sampling** *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*
IF a PRO-PM, identify whether (and how) proxy responses are allowed.
Not Applicable

S.21. **Survey/Patient-reported data** *(If measure is based on a survey, provide instructions for conducting the survey and guidance on minimum response rate.)*
IF a PRO-PM, specify calculation of response rates to be reported with performance measure results.
Not Applicable

S.22. **Missing data** (specify how missing data are handled, e.g., imputation, delete case.)
Required for Composites and PRO-PMs.
eMeasures are calculated using only the structured data collected in certified EHR technology (CEHRT). Data not present in the structured field from which the measure draws will not be included in the measure calculation.

S.23. **Data Source** *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*
If other, please describe in S.24.
Electronic Clinical Data, Electronic Clinical Data : Electronic Health Record

S.24. **Data Source or Collection Instrument** *(Identify the specific data source/data collection instrument e.g. name of database, clinical registry, collection instrument, etc.)*
IF a PRO-PM, identify the specific PROM(s); and standard methods, modes, and languages of administration.
Hospitals report EHR data using Certified Electronic Health Record Technology (CEHRT), and by submitting Quality Reporting Document Architecture Category 1 (QRDA-1).

S.25. **Data Source or Collection Instrument** *(available at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
No data collection instrument provided

S.26. **Level of Analysis** *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*
Facility, Population : National

S.27. **Care Setting** *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*
Hospital/Acute Care Facility

If other:

S.28. **COMPOSITE Performance Measure** - Additional Specifications *(Use this section as needed for aggregation and weighting rules, or calculation of individual performance measures if not individually endorsed.)*
Not Applicable

2a. Reliability – See attached Measure Testing Submission Form
2b. Validity – See attached Measure Testing Submission Form
PC-05_eCQM_testing_attachment.docx,NQF2830_CMS9v4_PC05_Bonnie_Testing-635908830409016436.xlsx

NATIONAL QUALITY FORUM—Measure Testing (subcriteria 2a2, 2b2-2b7)

Measure Number (*if previously endorsed*)**:** 2830

Measure Title: PC-05: Exclusive Breastmilk Feeding

Date of Submission: 2/11/2016

Type of Measure:

☐ Composite – ***STOP – use composite testing form***	☐ Outcome (*including PRO-PM*)
☐ Cost/resource	X Process
☐ Efficiency	☐ Structure

Instructions

- Measures must be tested for all the data sources and levels of analyses that are specified. ***If there is more than one set of data specifications or more than one level of analysis, contact NQF staff*** about how to present all the testing information in one form.
- **For all measures, sections 1, 2a2, 2b2, 2b3, and 2b5 must be completed.**
- **For outcome and resource use measures**, section **2b4** also must be completed.
- If specified for **multiple data sources/sets of specificaitons** (e.g., claims and EHRs), section **2b6** also must be completed.
- Respond to all questions as instructed with answers immediately following the question. All information on testing to demonstrate meeting the sub criteria for reliability (2a2) and validity (2b2-2b6) must be in this form. An appendix for *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 20 pages (*incuding questions/instructions;* minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.
- For information on the most updated guidance on how to address sociodemographic variables and testing in this form refer to the release notes for version 6.6 of the Measure Testing Attachment.

Note: The information provided in this form is intended to aid the Steering Committee and other stakeholders in understanding to what degree the testing results for this measure meet NQF's evaluation criteria for testing.

2a2. Reliability testing [10] demonstrates the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise. For **PRO-PMs and composite performance measures**, reliability should be demonstrated for the computed performance score.

2b2. Validity testing [11] demonstrates that the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality. For **PRO-PMs and composite performance measures**, validity should be demonstrated for the computed performance score.

2b3. Exclusions are supported by the clinical evidence; otherwise, they are supported by evidence of sufficient frequency of occurrence so that results are distorted without the exclusion; [12]
AND
If patient preference (e.g., informed decision-making) is a basis for exclusion, there must be evidence that the exclusion impacts performance on the measure; in such cases, the measure must be specified so that the information about patient preference and the effect on the measure is transparent (e.g., numerator category computed separately, denominator exclusion category computed separately). [13]

2b4. For outcome measures and other measures when indicated (e.g., resource use):

- **an evidence-based risk-adjustment strategy** (e.g., risk models, risk stratification) is specified; is based on patient factors (including clinical and sociodemographic factors) that influence the measured outcome and are present at start of care; [14,15] and has demonstrated adequate discrimination and calibration

OR

- rationale/data support no risk adjustment/ stratification.

2b5. Data analysis of computed measure scores demonstrates that methods for scoring and analysis of the specified measure allow for **identification of statistically significant and practically/clinically meaningful [16] differences in performance**;

OR

there is evidence of overall less-than-optimal performance.

2b6. If multiple data sources/methods are specified, there is demonstration they produce comparable results.

2b7. For **eMeasures, composites, and PRO-PMs** (or other measures susceptible to missing data), analyses identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and non-responders) and how the specified handling of missing data minimizes bias.

Notes

10. Reliability testing applies to both the data elements and computed measure score. Examples of reliability testing for data elements include, but are not limited to: inter-rater/abstractor or intra-rater/abstractor studies; internal consistency for multi-item scales; test-retest for survey items. Reliability testing of the measure score addresses precision of measurement (e.g., signal-to-noise).

11. Validity testing applies to both the data elements and computed measure score. Validity testing of data elements typically analyzes agreement with another authoritative source of the same information. Examples of validity testing of the measure score include, but are not limited to: testing hypotheses that the measures scores indicate quality of care, e.g., measure scores are different for groups known to have differences in quality assessed by another valid quality measure or method; correlation of measure scores with another valid indicator of quality for the specific topic; or relationship to conceptually related measures (e.g., scores on process measures to scores on outcome measures). Face validity of the measure score as a quality indicator may be adequate if accomplished through a systematic and transparent process, by identified experts, and explicitly addresses whether performance scores resulting from the measure as specified can be used to distinguish good from poor quality.

12. Examples of evidence that an exclusion distorts measure results include, but are not limited to: frequency of occurrence, variability of exclusions across providers, and sensitivity analyses with and without the exclusion.

13. Patient preference is not a clinical exception to eligibility and can be influenced by provider interventions.

14. Risk factors that influence outcomes should not be specified as exclusions

15. With large enough sample sizes, small differences that are statistically significant may or may not be practically or clinically meaningful. The substantive question may be, for example, whether a statistically significant difference of one percentage point in the percentage of patients who received smoking cessation counseling (e.g., 74 percent v. 75 percent) is clinically meaningful; or whether a statistically significant difference of $25 in cost for an episode of care (e.g., $5,000 v. $5,025) is practically meaningful. Measures with overall less-than-optimal performance may not demonstrate much variability across providers.

1. DATA/SAMPLE USED FOR ALL TESTING OF THIS MEASURE

Often the same data are used for all aspects of measure testing. In an effort to eliminate duplication, the first five questions apply to all measure testing. If there are differences by aspect of testing,(e.g., reliability vs. validity) be sure to indicate the specific differences in question 1.7.

1.1. What type of data was used for testing? (*Check all the sources of data identified in the measure specifications and data used for testing the measure*. *Testing must be provided for all the sources of data specified and intended for measure implementation.* ***If different data sources are used for the numerator and denominator, indicate N [numerator] or D [denominator] after the checkbox.***)

Measure Specified to Use Data From: (*must be consistent with data sources entered in S.23*)	Measure Tested with Data From:
☐ abstracted from paper record	☐ abstracted from paper record
☐ administrative claims	☐ administrative claims
☐ clinical database/registry	☐ clinical database/registry
☐ abstracted from electronic health record	☐ abstracted from electronic health record
☒ eMeasure (HQMF) implemented in EHRs	☐ eMeasure (HQMF) implemented in EHRs
☒other: Bonnie Test Cases	☒ other: Bonnie Test Cases (see question 1.2)

1.2. If an existing dataset was used, identify the specific dataset (*the dataset used for testing must be consistent with the measure specifications for target population and healthcare entities being measured; e.g., Medicare Part A claims, Medicaid claims, other commercial insurance, nursing home MDS, home health OASIS, clinical registry*).

This measure is a legacy eCQM (an NQF endorsed measure that has been re-specified into an eMeasure and is currently used in a federal quality program/programs). In accordance with the modified NQF requirements for testing of eMeasures, the Bonnie testing tool was used to simulate a testing environment where measure specifications and HQMF output are tested against synthetic test data. Measure developers rely on the results in Bonnie to confirm whether the measure logic is performing as expected. Reference the eCQI Resource Center website (https://ecqi.healthit.gov/ecqm-tools/tool-library/bonnie) or the Bonnie testing tool website (https://bonnie.healthit.gov/) for more information about Bonnie functionality and its role in measure development. Please also reference the Bonnie testing worksheet attachment for detailed Bonnie test cases and testing results for this measure.

While Bonnie does consume HQMF specifications, we are not able to test the measure with data from HQMF implemented in EHRs.

1.3. What are the dates of the data used in testing? The measure specifications were tested in the Bonnie testing environment which mimics the year 2012.

1.4. What levels of analysis were tested? (*testing must be provided for all the levels specified and intended for measure implementation, e.g., individual clinician, hospital, health plan*)

Measure Specified to Measure Performance of: (*must be consistent with levels entered in item S.26*)	Measure Tested at Level of:
☐ individual clinician	☐ individual clinician
☐ group/practice	☐ group/practice
☒ hospital/facility/agency	☒ hospital/facility/agency
☐ health plan	☐ health plan
☐ other:	☒ other: Bonnie Test Cases

1.5. How many and which measured entities were included in the testing and analysis (by level of analysis and data source)? (*Identify the number and descriptive characteristics of measured entities included in the analysis (e.g., size, location, type); if a sample was used, describe how entities were selected for inclusion in the sample*)

Not applicable.

1.6. How many and which patients were included in the testing and analysis (by level of analysis and data source)? (*Identify the number and descriptive characteristics of patients included in the analysis (e.g., age, sex, race, diagnosis); if a sample was used, describe how patients were selected for inclusion in the sample*)

28 unique synthetic patient records were created in the BONNIE testing system for this measure. Cases were used to test the validity of each data element and timing relationship in the measure. Patient characteristics such as gestational

age, diagnosis, and length of stay were pre-determined to provide a variety of scenarios that adequately tested patients passing each data element and failing each data element. Data included in cases and tested for this measure included diagnoses, gestational ages, forms of nutrition such as breast milk and parenteral nutrition, discharge statuses, and level of care.

All 28 cases passed or failed as expected based on the data included in the case, confirming the measure logic is accurate and valid.

1.7. If there are differences in the data or sample used for different aspects of testing (e.g., reliability, validity, exclusions, risk adjustment), identify how the data or sample are different for each aspect of testing reported below.

Not applicable. Bonnie test patients were developed for use across the different aspects of testing.

1.8 What were the patient-level sociodemographic (SDS) variables that were available and analyzed in the data or sample used? For example, patient-reported data (e.g., income, education, language), proxy variables when SDS data are not collected from each patient (e.g. census tract), or patient community characteristics (e.g. percent vacant housing, crime rate).

Not applicable.

2a2. RELIABILITY TESTING

Note*: If accuracy/correctness (validity) of data elements was empirically tested, separate reliability testing of data elements is not required – in 2a2.1 check critical data elements; in 2a2.2 enter "see section 2b2 for validity testing of data elements"; and skip 2a2.3 and 2a2.4.*

2a2.1. What level of reliability testing was conducted? (*May be one or both levels*)

☐ **Critical data elements used in the measure** (*e.g., inter-abstractor reliability; data element reliability must address ALL critical data elements*)
☐ **Performance measure score** (e.g., *signal-to-noise analysis*)

Not applicable. Synthetic test patients were created in the Bonnie testing environment.

2a2.2. For each level checked above, describe the method of reliability testing and what it tests (*describe the steps—do not just name a method; what type of error does it test; what statistical analysis was used*)

Not applicable. Synthetic test patients were created in the Bonnie testing environment.

2a2.3. For each level of testing checked above, what were the statistical results from reliability testing? (e*.g., percent agreement and kappa for the critical data elements; distribution of reliability statistics from a signal-to-noise analysis*)

Not applicable. Synthetic test patients were created in the Bonnie testing environment.

2a2.4 What is your interpretation of the results in terms of demonstrating reliability? (i.*e., what do the results mean and what are the norms for the test conducted?*)

Not applicable. Synthetic test patients were created in the Bonnie testing environment.

2b2. VALIDITY TESTING

2b2.1. What level of validity testing was conducted? (*may be one or both levels*)

☒ **Critical data elements** (*data element validity must address ALL critical data elements*)

☒ **Performance measure score**

☐ **Empirical validity testing**

☒ **Systematic assessment of face validity of performance measure score as an indicator** of quality or resource use (*i.e., is an accurate reflection of performance on quality or resource use and can distinguish good from poor performance*)

2b2.2. For each level of testing checked above, describe the method of validity testing and what it tests (*describe the steps—do not just name a method; what was tested, e.g., accuracy of data elements compared to authoritative source, relationship to another measure as expected; what statistical analysis was used)*

The Bonnie testing tool and environment were used to test the measure logic and value sets. Each data element and logic statement was tested to confirm actual results met expectations. Bonnie testing includes negative and positive testing of each data element in the measure. Positive testing ensures patients expected to be included in the measure are included. Negative testing ensures that patients who do not meet the data criteria are not included in the measure. An example of negative testing would be to include test cases with pediatric ages to ensure that pediatric patients are not included in the measure.

Denominator test cases positively test to ensure singleton newborns with a gestational age >=37 weeks and who do not have galactosemia or parenteral nutrition are included in the denominator. Positive test cases include patients with a diagnosis of galactosemia or an order for parenteral nutrition, to ensure these cases are appropriately removed from the measure. Negative test cases ensure patients who do not meet denominator criteria, such as patients with a gestational age <37 weeks, fall out of the denominator population.

Numerator test cases positively test to ensure patients who are exclusively fed breastmilk fall in to the measure. Negative test cases ensure that patients who do not exclusively receive breastmilk, or those without documentation of exclusively receiving breastmilk, are not included in the numerator.

Denominator exclusion and exception test cases for this measure ensure that patients are properly removed from the denominator if they have specific discharge statuses other than discharged home, or if they require a NICU level of care. Negative test cases are also run. For example, patients who do not have a NICU level of care are expected to remain in the denominator, rather than falling in to the exclusion. Testing confirmed patients meeting the exclusion and exception criteria are removed from the measure appropriately, while those that do not meet the criteria are retained in the denominator population.

A review of the measure specifications was also conducted to confirm the logic was properly expressed within the current version of the QDM and confirmed the logic matches the clinical intent of the measure, as stated in the measure header. This is done through use of a logic checklist to facilitate review of the measure logic, according to several checks, a few examples of which are included below:

- Is the intent of the measure described in the measure description articulated/ captured in the measure logic?
- Do the logic elements map to definitions in the measure narrative, data dictionary or supporting reference documentation?
- Do the populations in the narrative align with the populations defined in the logic?
- Are the mathematic inequalities reflective of the measure intent and represent the intended populations (for example: when intended the inequality represents less than rather than less and or equal to)?

2b2.3. What were the statistical results from validity testing? (*e.g., correlation; t-test*)

Bonnie results provide coverage and passing rates. This measure reached 100% coverage, confirming there is a test case for each pathway of logic. This measure also has a 100% passing rate, confirming all test cases performed as expected.

2b2.4. What is your interpretation of the results in terms of demonstrating validity? (i.*e., what do the results mean and what are the norms for the test conducted?*)

This measure performed as expected in Bonnie.

2b3. EXCLUSIONS ANALYSIS

NA ☐ no exclusions — *skip to section 2b4*

2b3.1. Describe the method of testing exclusions and what it tests (*describe the steps—do not just name a method; what was tested, e.g., whether exclusions affect overall performance scores; what statistical analysis was used*)

Exclusions in the eMeasure align with the chart-based version of the measure, and are clinically necessary for the interpretation of the measure. No additional exclusion analyses were done on the eMeasure.

2b3.2. What were the statistical results from testing exclusions? (*include overall number and percentage of individuals excluded, frequency distribution of exclusions across measured entities, and impact on performance measure scores*)

Not applicable.

2b3.3. What is your interpretation of the results in terms of demonstrating that exclusions are needed to prevent unfair distortion of performance results? (*i.e., the value outweighs the burden of increased data collection and analysis. Note: **If patient preference is an exclusion**, the measure must be specified so that the effect on the performance score is transparent, e.g., scores with and without exclusion*)

Not applicable.

2b4. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES

If not an intermediate or health outcome, or PRO-PM, or resource use measure, skip to section 2b5.

2b4.1. What method of controlling for differences in case mix is used?

- ☒ **No risk adjustment or stratification**
- ☐ **Statistical risk model with** Click here to enter number of factors **risk factors**
- ☐ **Stratification by** Click here to enter number of categories **risk categories**
- ☐ **Other,** Click here to enter description

2b4.2. If an outcome or resource use measure is not risk adjusted or stratified, provide rationale and analyses to demonstrate that controlling for differences in patient characteristics (case mix) is not needed to achieve fair comparisons across measured entities.

2b4.3. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors (clinical factors or sociodemographic factors) used in the statistical risk model or for stratification by risk (*e.g., potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of p<0.10; correlation of x or higher; patient factors should be present at the start of care*)

2b4.4a. What were the statistical results of the analyses used to select risk factors?

2b4.4b. Describe the analyses and interpretation resulting in the decision to select SDS factors (e.g. prevalence of the factor across measured entities, empirical association with the outcome, contribution of unique variation in the outcome, assessment of between-unit effects and within-unit effects)

2b4.5. Describe the method of testing/analysis used to develop and validate the adequacy of the statistical model or stratification approach (*describe the steps—do not just name a method; what statistical analysis was used*)

Provide the statistical results from testing the approach to controlling for differences in patient characteristics (case mix) below.
If stratified, skip to 2b4.9

2b4.6. Statistical Risk Model Discrimination Statistics (*e.g., c-statistic, R-squared*)**:**

2b4.7. Statistical Risk Model Calibration Statistics (*e.g., Hosmer-Lemeshow statistic*):

2b4.8. Statistical Risk Model Calibration – Risk decile plots or calibration curves:

2b4.9. Results of Risk Stratification Analysis:

2b4.10. What is your interpretation of the results in terms of demonstrating adequacy of controlling for differences in patient characteristics (case mix)? (i.*e., what do the results mean and what are the norms for the test conducted*)

2b4.11. Optional Additional Testing for Risk Adjustment (*not required, but would provide additional support of adequacy of risk model, e.g., testing of risk model in another data set; sensitivity analysis for missing data; other methods that were assessed*)

2b5. IDENTIFICATION OF STATISTICALLY SIGNIFICANT & MEANINGFUL DIFFERENCES IN PERFORMANCE

2b5.1. Describe the method for determining if statistically significant and clinically/practically meaningful differences in performance measure scores among the measured entities can be identified (*describe the steps—do not just name a method; what statistical analysis was used? Do not just repeat the information provided related to performance gap in 1b)*

Not applicable.

2b5.2. What were the statistical results from testing the ability to identify statistically significant and/or clinically/practically meaningful differences in performance measure scores across measured entities? (e.g., *number and percentage of entities with scores that were statistically significantly different from mean or some benchmark, different from expected; how was meaningful difference defined*)

Not applicable.

2b5.3. What is your interpretation of the results in terms of demonstrating the ability to identify statistically significant and/or clinically/practically meaningful differences in performance across measured entities? (i.*e., what do the results mean in terms of statistical and meaningful differences?*)

Not applicable.

2b6. COMPARABILITY OF PERFORMANCE SCORES WHEN MORE THAN ONE SET OF SPECIFICATIONS
***If only one set of specifications, this section can be skipped*.**

Note: *This item is directed to measures that are risk-adjusted (with or without SDS factors)* ***OR*** *to measures with more than one set of specifications/instructions (e.g., one set of specifications for how to identify and compute the measure from medical record abstraction and a different set of specifications for claims or eMeasures). It does not apply to measures that use more than one source of data in one set of specifications/instructions (e.g., claims data to identify the denominator and medical record abstraction for the numerator).* ***Comparability is not required when comparing performance scores with and without SDS factors in the risk adjustment model. However, if comparability is not demonstrated for measures with more than one set of specifications/instructions, the different specifications (e.g., for medical records vs. claims) should be submitted as separate measures.***

2b6.1. Describe the method of testing conducted to compare performance scores for the same entities across the different data sources/specifications (*describe the steps—do not just name a method; what statistical analysis was used*)

Not applicable.

2b6.2. What were the statistical results from testing comparability of performance scores for the same entities when using different data sources/specifications? (*e.g., correlation, rank order*)

Not applicable.

2b6.3. What is your interpretation of the results in terms of the differences in performance measure scores for the same entities across the different data sources/specifications? (i.*e., what do the results mean and what are the norms for the test conducted*)

Not applicable.

2b7. MISSING DATA ANALYSIS AND MINIMIZING BIAS

2b7.1. Describe the method of testing conducted to identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias (*describe the steps—do not just name a method; what statistical analysis was used*)

Data not present in the structured field from which the measure draws will not be included in the measure calculation. In the Bonnie testing environment, missing data are tested as an expected "Fail" for that data element, and actual performance (whether or not the case fails) is compared to expectations to ensure missing data impacts data element and overall measure calculation as expected. This is the extent of missing data analysis that can be performed in Bonnie.

2b7.2. What is the overall frequency of missing data, the distribution of missing data across providers, and the results from testing related to missing data? (*e.g., results of sensitivity analysis of the effect of various rules for missing data/nonresponse; if no empirical sensitivity analysis, identify the approaches for handling missing data that were considered and pros and cons of each*)

See response for question 2b7.1 above.

2b7.3. What is your interpretation of the results in terms of demonstrating that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias**?** (i.*e., what do the results mean in terms of supporting the selected approach for missing data and what are the norms for the test conducted; if no empirical analysis, provide rationale for the selected approach for missing data*)

The measure performed as expected in the Bonnie testing environment.

3. Feasibility

Extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

3a. Byproduct of Care Processes

For clinical measures, the required data elements are routinely generated and used during care delivery (e.g., blood pressure, lab test, diagnosis, medication order).

3a.1. **Data Elements Generated as Byproduct of Care Processes.**

Generated or collected by and used by healthcare personnel during the provision of care (e.g., blood pressure, lab value, diagnosis, depression score)

If other:

3b. Electronic Sources

The required data elements are available in electronic health records or other electronic sources. If the required data are not in electronic health records or existing electronic sources, a credible, near-term path to electronic collection is specified.

3b.1. **To what extent are the specified data elements available electronically in defined fields?** *(i.e., data elements that are needed to compute the performance measure score are in defined, computer-readable fields)*

ALL data elements are in defined fields in electronic health records (EHRs)

3b.2. **If ALL the data elements needed to compute the performance measure score are not from electronic sources, specify a credible, near-term path to electronic capture, OR provide a rationale for using other than electronic sources.**

3b.3. **If this is an eMeasure, provide a summary of the feasibility assessment in an attached file or make available at a measure-specific URL.**

Attachment **Attachment:** PC05_eCQM_NQF_Measure_Feasibility_Assessment_Report.docx

3c. Data Collection Strategy

Demonstration that the data collection strategy (e.g., source, timing, frequency, sampling, patient confidentiality, costs associated with fees/licensing of proprietary measures) can be implemented (e.g., already in operational use, or testing demonstrates that it is ready to put into operational use). For eMeasures, a feasibility assessment addresses the data elements and measure logic and demonstrates the eMeasure can be implemented or feasibility concerns can be adequately addressed.

3c.1. **Describe what you have learned/modified as a result of testing and/or operational use of the measure regarding data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, time and cost of data collection, other feasibility/implementation issues.**

IF a PRO-PM, consider implications for both individuals providing PROM data (patients, service recipients, respondents) and those whose performance is being measured.

Not Applicable

3c.2. **Describe any fees, licensing, or other requirements to use any aspect of the measure as specified** *(e.g., value/code set, risk model, programming code, algorithm)*.

Value sets are housed in the Value Set Authority Center (VSAC), which is provided by the National Library of Medicine (NLM), in coordination with the Office of the National Coordinator for Health Information Technology and the Centers for Medicare & Medicaid Services.

Viewing or downloading value sets requires a free Unified Medical Language System® (UMLS) Metathesaurus License, due to usage restrictions on some of the codes included in the value sets. Individuals interested in accessing value set content can request a UMLS license at https://uts.nlm.nih.gov/license.html)

There are no other fees or licensing requirements to use the Joint Commission performance measures, all of which are in the public domain.

4. Usability and Use

Extent to which potential audiences (e.g., consumers, purchasers, providers, policy makers) are using or could use performance results for both accountability and performance improvement to achieve the goal of high-quality, efficient healthcare for individuals or populations.

4a. Accountability and Transparency

Performance results are used in at least one accountability application within three years after initial endorsement and are publicly reported within six years after initial endorsement (or the data on performance results are available). If not in use at the time of initial endorsement, then a credible plan for implementation within the specified timeframes is provided.

4.1. Current and Planned Use

NQF-endorsed measures are expected to be used in at least one accountability application within 3 years and publicly reported within 6 years of initial endorsement in addition to performance improvement.

Planned	Current Use (for current use provide URL)
Public Reporting Quality Improvement with Benchmarking (external benchmarking to multiple organizations) Quality Improvement (Internal to the specific organization)	Payment Program Hospital Inpatient Quality Reporting Program https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/HospitalQualityInits/HospitalRHQDAPU.html Regulatory and Accreditation Programs Hospital Accreditation Program http://jointcommission.org EHR Incentive Program https://www.cms.gov/Regulations-and-Guidance/Legislation/EHRIncentivePrograms/index.html?redirect=/ehrincentiveprograms

4a.1. For each CURRENT use, checked above, provide:

- Name of program and sponsor
- Purpose
- Geographic area and number and percentage of accountable entities and patients included

- Name of program and sponsor: Hospital Inpatient Quality Reporting Program; Centers for Medicare & Medicaid Services
- Purpose: The Hospital Inpatient Quality Reporting (Hospital IQR) program was originally mandated by Section 501(b) of the Medicare Prescription Drug, Improvement, and Modernization Act (MMA) of 2003. This section of the MMA authorized CMS to pay hospitals that successfully report designated quality measures a higher annual update to their payment rates.
- Geographic area and number and percentage of accountable entities and patients included: Nationwide; 3500 hospitals (2015)

- Name of program and sponsor: Hospital Accreditation Program; The Joint Commission
- Purpose: An accreditation program that recognizes hospitals that meet standard requirements to provide safe and effective care
- Geographic area and number an percentage of accountable entities and patients included; Nationwide; 3,300 Joint Commission-accredited hospitals (2014)

- Name of program and sponsor: EHR Incentive Program; Centers for Medicare & Medicaid Services
- Purpose: The American Recovery and Reinvestment Act of 2009 (ARRA) (Pub.L. 111–5) was enacted on February 17, 2009. Title IV of Division B of ARRA amends Titles XVIII and XIX of the Social Security Act (the Act) by establishing incentive payments to eligible professionals (EPs), eligible hospitals, and critical access hospitals (CAHs), and Medicare Advantage Organizations to promote the adoption and meaningful use of interoperable health information technology (HIT) and qualified electronic health records (EHRs). These incentive payments are part of a broader effort under the HITECH Act to accelerate the adoption of HIT and utilization of qualified EHRs.
- Geographic area and number and percentage of accountable entities and patients included: Nationwide; 4,847 active registrations (September, 2015)

4a.2. **If not currently publicly reported OR used in at least one other accountability application (e.g., payment program, certification, licensing) what are the reasons?** (*e.g., Do policies or actions of the developer/steward or accountable entities restrict access to performance results or impede implementation?*)
Not Applicable

4a.3. **If not currently publicly reported OR used in at least one other accountability application, provide a credible plan for implementation within the expected timeframes -- any accountability application within 3 years and publicly reported within 6 years of initial endorsement.** (*Credible plan includes the specific program, purpose, intended audience, and timeline for implementing the measure within the specified timeframes. A plan for accountability applications addresses mechanisms for data aggregation and reporting.*)
Not Applicable

4b. Improvement

Progress toward achieving the goal of high-quality, efficient healthcare for individuals or populations is demonstrated. If not in use for performance improvement at the time of initial endorsement, then a credible rationale describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.

4b.1. **Progress on Improvement. (Not required for initial endorsement unless available.)**
Performance results on this measure (current and over time) should be provided in 1b.2 and 1b.4. Discuss:

- **Progress (trends in performance results, number and percentage of people receiving high-quality healthcare)**
- **Geographic area and number and percentage of accountable entities and patients included**

While the chart-abstracted version of this measure has indicated improvement over time, this measure is a legacy eCQM that is currently in use in the Hospital Inpatient Quality Reporting Program (HIQR) and the EHR Incentive Program. At present, no performance data are currently available.

In CY2016, CMS is requiring organizations participating in HIQR to electronically submit 1 quarter of data on 4 of 28 available eCQMs, one of which is this measure. Data will be accepted through February 28th, 2017.

For more information, refer to CY2016 IPPS Final Rule, located here: https://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/AcuteInpatientPPS/FY2016-IPPS-Final-Rule-Home-Page.html

4b.2. **If no improvement was demonstrated, what are the reasons? If not in use for performance improvement at the time of initial endorsement, provide a credible rationale that describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.**
Not Applicable

4c. Unintended Consequences

The benefits of the performance measure in facilitating progress toward achieving high-quality, efficient healthcare for individuals or populations outweigh evidence of unintended negative consequences to individuals or populations (if such evidence exists).

4c.1. **Were any unintended negative consequences to individuals or populations identified during testing; OR has evidence of unintended negative consequences to individuals or populations been reported since implementation? If so, identify the negative unintended consequences and describe how benefits outweigh them or actions taken to mitigate them.**
Not Applicable

5. Comparison to Related or Competing Measures

If a measure meets the above criteria and there are endorsed or new related measures (either the same measure focus or the same target population) or competing measures (both the same measure focus and the same target population), the measures are compared to address harmonization and/or selection of the best measure.

5. Relation to Other NQF-endorsed Measures
Are there related measures (conceptually, either same measure focus or target population) or competing measures (conceptually both the same measure focus and same target population)? If yes, list the NQF # and title of all related and/or competing measures.
Yes

5.1a. **List of related or competing measures (selected from NQF-endorsed measures)**
0480 : PC-05 Exclusive Breast Milk Feeding

5.1b. **If related or competing measures are not NQF endorsed please indicate measure title and steward.**

5a. Harmonization

The measure specifications are harmonized with related measures;
OR
The differences in specifications are justified

5a.1. **If this measure conceptually addresses EITHER the same measure focus OR the same target population as NQF-endorsed measure(s):**
Are the measure specifications completely harmonized?
Yes

5a.2. **If the measure specifications are not completely harmonized, identify the differences, rationale, and impact on interpretability and data collection burden.**
#0480: Exclusive Breast Milk Feeding: The measures are completely harmonized to the extent possible, given the fact that the data source for #0480 is the paper medical record, and the data source for #2830 is the electronic health record.

5b. Competing Measures

The measure is superior to competing measures (e.g., is a more valid or efficient way to measure);
OR
Multiple measures are justified.

5b.1. **If this measure conceptually addresses both the same measure focus and the same target population as NQF-endorsed measure(s):**
Describe why this measure is superior to competing measures (e.g., a more valid or efficient way to measure quality); OR provide a rationale for the additive value of endorsing an additional measure. (Provide analyses when possible.)
Not Applicable

Appendix

A.1 Supplemental materials may be provided in an appendix. All supplemental materials (such as data collection instrument or methodology reports) should be organized in one file with a table of contents or bookmarks. If material pertains to a specific submission form number, that should be indicated. Requested information should be provided in the submission form and required attachments. There is no guarantee that supplemental materials will be reviewed.

Attachment:

Contact Information

Co.1 **Measure Steward (Intellectual Property Owner):** The Joint Commission

Co.2 **Point of Contact:** Ann, Watt, awatt@jointcommission.org, 630-792-5944-
Co.3 **Measure Developer if different from Measure Steward:** The Joint Commission
Co.4 **Point of Contact:** Michelle, Dardis, mdardis@jointcommission.org, 630-792-5066-

Additional Information

Ad.1 **Workgroup/Expert Panel involved in measure development**
Provide a list of sponsoring organizations and workgroup/panel members' names and organizations. Describe the members' role in measure development.
The technical advisory panel (TAP) members determined priority areas that could be evaluated to improve care related to perinatal care during the development timeframe. After implementation, minor revisions, acknowledged by TAP representatives, were made to improve clarity. Hospital feedback will be reviewed during the reliability testing phase of the project to assist the TAP in making the final measure recommendations. Members are:

Michael Ross, MD, MPH (Chair)
Harbor-UCLA Medical Center
Torrance, CA

Wanda Barfield, MD, MPH
Centers for Disease Control and Prevention
Atlanta, GA

Kenneth E. Brown, MD, MBA, FACOG, FACHE
Woman´s Hospital
Lafayette, LA

Martin McCaffrey, MD
UNC North Carolina Children's Hospital
Chapel Hill, NC

Cathy Collins-Fulea, MSN, CNM
Henry Ford Hospital
Detroit, MI

Janet H. Muri, MBA
National Perinatal Information Center/
Quality Analytic Services
Providence, RI

Kathleen Simpson, PhD, RNC, FAAN
St. John's Mercy Medical Center
St. Louis, MO

Michael Socol, MD
Northwestern University Medical School
Chicago, IL

Rebecca Zimmermann, MPP
America's Health Insurance Plans
Washington, DC

Measure Developer/Steward Updates and Ongoing Maintenance
Ad.2 **Year the measure was first released:** 2012
Ad.3 **Month and Year of most recent revision:** 05, 2015
Ad.4 **What is your frequency for review/update of this measure?** Annual
Ad.5 **When is the next scheduled review/update for this measure?** 04, 2015

Ad.7 **Disclaimers:** These performance measures are not clinical guidelines and do not establish a standard of medical care, and have not been tested for all potential applications. The measures and specifications are provided without warranty.

Ad.8 **Additional Information/Comments:** Release notes are published each year to reflect changes made in the annual update of the electronic specifications. For changes made in the most recent update, please refer to the "2014 eCQM Technical Release Notes Update June 2015" document, found at: https://www.cms.gov/Regulations-and-Guidance/Legislation/EHRIncentivePrograms/Downloads/EHandEPTRNs.pdf

NATIONAL QUALITY FORUM

MEASURE WORKSHEET

This document summarizes the evaluation of the measure as it progresses through NQF's Consensus Development Process (CDP). The information submitted by measure developers/stewards is included after the Brief Measure Information, Preliminary Analysis, and Pre-meeting Public and Member Comments sections.

To navigate the links in the worksheet: Ctrl + click link to go to the link; ALT + LEFT ARROW to return

Brief Measure Information

NQF #: 2892
Measure Title: Birthrisk Cesarean Birth Measure
Measure Steward: Birthrisk.com, LLC.
Brief Description of Measure: This is a measure of the effect that obstetrical care provider´s labor management strategies have on their laboring patient´s risk for cesarean birth. The target population is limited to women who attempt labor with a singleton vertex pregnancy without a history of a prior cesarean birth and give birth between 37 and 42 weeks of gestation.
Developer Rationale: Healthcare administrators, hospital administrators and obstetrical care providers do not currently have an accurate method for determining if their labor management strategies are better or worse than the average. The reason that an accurate method is not available is because the physical process of labor makes it impossible to assess the effect of the obstetrical care provider´s labor management strategies unless the unique combination of physical characteristics of the mother and the size of her baby are taken into account.

Empowering obstetrical care providers with an accurate cesarean birth measure will allow them to analyze which labor management strategies work best at reaching the appropriate cesarean birth rate for their population of patients. Hospital administrators will have the ability to conduct and monitor quality improvement projects aimed at achieving the appropriate cesarean birth rate for their institution by analyzing the results achieved by each obstetrical care provider in their hospital. Healthcare administrators will be able to monitor the various organizational units under their jurisdiction in order to identify centers of excellence as well as centers that need improvement. With the ability to compare organizational units to the average in the state or nation, healthcare administrators could initiate a state/nation wide incentive/disincentive program for entities that are statistically significantly better/worse than the average in the state/nation.

Numerator Statement: Number of cesarean births.
Denominator Statement: Women without a history of a prior cesarean birth who attempted labor and gave birth to a single baby in vertex presentation between 37 and 42 weeks of gestation.
Denominator Exclusions: The denominator excludes women with any of the following:

1. Gestational age at birth of less than 37 weeks or greater than 42 weeks.
2. History of a prior cesarean birth.
3. Multiple gestation.
4. Not in vertex presentation.
5. Did not attempt to have a vaginal birth by attempting labor.

Measure Type: Outcome
Data Source: Other
Level of Analysis: Clinician : Individual, Facility

IF Endorsement Maintenance – Original Endorsement Date: Most Recent Endorsement Date:

IF this measure is included in a composite, NQF Composite#/title:

IF this measure is paired/grouped, NQF#/title:

IF PAIRED/GROUPED, what is the reason this measure must be reported with other measures to appropriately interpret results?
N/A

New Measure -- Preliminary Analysis

Criteria 1: Importance to Measure and Report

1a. Evidence

1a. Evidence. The evidence requirements for a health outcomes measure include providing rationale that supports the relationship of the health outcome to processes or structures of care. The guidance for evaluating the clinical evidence asks if the relationship between the measured health outcome and at least one clinical action is identified and supported by the stated rationale.

Summary of evidence:

- The developer states "There are many different labor management strategies that have been used over the years to assist women who are in labor. Some of these strategies can decrease a woman's inherent risk that labor will result in a cesarean birth and others can increase her inherent risk. For example, the use of forceps can decrease a woman's inherent risk for a cesarean birth whereas an impatient obstetrical care provider may increase a woman's inherent risk."

Question for the Committee:

- *Is there at least one thing that the provider can do to achieve a change in the measure results?*

Preliminary rating for evidence: ☒ **Pass** ☐ **No Pass**

1b. Gap in Care/Opportunity for Improvement and 1b. disparities

1b. Performance Gap. The performance gap requirements include demonstrating quality problems and opportunity for improvement.

The developer reports on data from 2005 – 2007 from New York:

Level of analysis	#births	Mean (SD)	Min - Max	IQR
103 facilities	211,374	15.57% (4.52)	7% - 29%	5%
892 providers	209,300	14.31% (11.07)	0% - 71%	17%

Disparities

- The developer referenced a study (Ehrenthal et al. 2010)a retrospective cohort study of cesarean births among nulliparous women delivering a live, singleton, vertex pregnancy at term using clinical data from electronic hospital obstetric records at a large, regional, obstetric hospital, approximating a population-based cohort. The results showed a greater odds of cesarean birth from the following sociodemographic characteristics: black race, marital status, patient type (private or service), insurance type (Medicaid or private), and age older than 35 years.
- Another referenced study (Declercq et al.2015) used a sample of 2,233,144 women who had a singleton, vertex, term (37–41 weeks) birth in 2012 and no prior cesarean to demonstrate that prepregnancy obesity was found to represent an independent risk factor for primary cesareans. Their results showed that obesity rates were highest among American Indian and Alaska Native (32.5%) and non-Hispanic black mothers (30.5%).

Questions for the Committee:

- *Is there a gap in care that warrants a national performance measure?*
- *Is this measure useful in understanding disparities in this area of healthcare?*

Preliminary rating for opportunity for improvement: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 1: Importance to Measure and Report (including 1a, 1b, 1c)

1a. Evidence to Support Measure Focus

Comments:

Cesarean is a process measure but does not in and of itself say whether the labor was well managed or not which is the purported purpose of the measure.

Evidence is directly related to the outcome. Consistent and evidence based ways to manage labor need to be implemented nationally for nulliparous and multiparous patients.

1b. Performance Gap

Comments:

Variation here does not necessarily mean performance gap.

Performance data was provided. There are greater odds of a cesarean delivery with certain sociodemographic characteristics and data provided regarding age, race, and payer mix.

Multiple elements of the measure are known to be unreliable in birth certificated data. ie was there a previous cesarean, is the labor induced, was there an attempt at labor. Additionally, many of the risk adjustment variables have a high likelihood of being missing (maternal height for example). All other previous literature shows R values lower than the .98 cited here. I am concerned that there was not a separate development sample from the predicted data set, leading to very high rates. Not enough detail is given about the development of the model to determine this.

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability

2a1. Reliability Specifications

2a1. Specifications requires the measure, as specified, to produce consistent (reliable) and credible (valid) results about the quality of care when implemented.

Data source(s): Birth certificates

Specifications:

- The level of analysis is individual clinician or facility.
- Numerator is the number of women undergoing Cesarean births .
- The denominator is all women without a history of a prior cesarean birth who attempted labor and gave birth to a single baby in vertex presentation between 36 weeks 4 days and 42 weeks 3 days of gestation during the specified time period.
- Exclusions include: gestational age at birth of less than 36 weeks 4 days or greater than 42 weeks and 3 days; History of a prior cesarean birth; Multiple gestation; Not in vertex presentation; Did not attempt to have a vaginal birth by attempting labor.
- Risk-adjustment by cohort comparison. “The statistical risk model uses an unprecedented method (cohort comparison to previously recorded births) to determine the expected cesarean birth rate for the target population.” Discussed further under Risk adjustment.

Questions for the Committee :

- *Are all the data elements clearly defined? Are all appropriate codes included?*
- *Is the logic or calculation algorithm clear?*
- *How will a hospital use this measure?*
- *Is it likely this measure can be consistently implemented?*

2a2. Reliability Testing Testing attachment

2a2. Reliability testing demonstrates if the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise enough to distinguish differences in performance across providers.

SUMMARY OF TESTING

Reliability testing level ☒ **Measure score** ☐ **Data element** ☐ **Both**

Reliability testing performed with the data source and level of analysis indicated for this measure ☒ **Yes** ☐ **No**

Method(s) of reliability testing

- Characteristics of the data set are described in section 1.6.
- The developer reports on "A signal-to-noise analysis of the hospital and provider measure scores using ANOVA "was conducted using data from 2005-2007.
- The developer reports of data element validity testing of most, but not all birth certificate data elements in the validity section -- data element validity also counts for data element reliability.

Results of reliability testing

- The developer reports the F statistic for ANOVA to be 0.091 (p=0.913) for hospitals and 1.096 (p=0.334) for providers. The developer concludes "A ratio higher than 1:1 (greater than 0 dB) indicates more signal than noise. Thus, the between-group variance (signal) is comparable to the within-group variance. The Cronbach's alpha was .87 that was greater than .70 ("rule of thumb" Nunnally, 1978) indicating high reliability."
- NQF's statistical advisors are reviewing this data to provide an interpretation of the ANOVA results for the in-person meeting.

Guidance from the Reliability Algorithm

Specifications precise (Box 1) →empirical reliability testing of data elements →see validity

Questions for the Committee:

- *Is 2005 – 2007 data an appropriate dataset to test this measure?*
- *Is the test sample adequate to generalize for widespread implementation?*
- *Is the method used to compute the signal-to-noise reliability statistic appropriate?*
- *Do the results demonstrate sufficient reliability so that differences in performance can be identified?*

Preliminary rating for reliability: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

The rating based on data element validity testing is also used for reliability.

2b. Validity

2b1. Validity: Specifications

2b1. Validity Specifications. This section should determine if the measure specifications are consistent with the evidence.

Specifications consistent with evidence in 1a. ☒ **Yes** ☐ **Somewhat** ☐ **No**

Question for the Committee:

- *Are the specifications consistent with the evidence?*

2b2. Validity testing

2b2. Validity Testing should demonstrate the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality.

SUMMARY OF TESTING

Validity testing level ☐ **Measure score** ☐ **Data element testing against a gold standard** ☒ **Both**

Method of validity testing of the measure score:

- ☐ **Face validity only**
- ☒ **Empirical validity testing of the measure score**

Validity testing method:

- Data element: The developer provides data element validity on many of the data elements, but not all, from a 2013 CDC report on accuracy of birth certificate data evaluated in two states.
- Measure score: The developer reported that "Performance measure score validity testing was conducted using construct validity for the hospital performance measure score and criterion validity for the provider performance measure score. The type of criterion validity that was used was predictive validity."

Validity testing results:

- Data element testing:
  - The developer reports "The data elements that have been studied were demonstrated to have high validity when compared to the medical record with the exception of Induction of Labor and Trial of labor which were deemed to have substantial validity. There was a discrepancy in Obstetric estimate of gestation at delivery (exact weeks) between the two states with state B only having moderate validity. "

  however "Unfortunately not all of the data elements that are relevant to this measure have been tested for validity. The data elements of Mother's Height, Mother's Prepregnancy Weight and Mother's Weight at Delivery were not addressed any in of the studies. "
- Score-level testing:
  - The developer describes the methodology used for score-level validation as construct validity for the hospital-level measure and criterion (specifically, predictive) validity for the provider-level measure. Both of these methods require a premise about the relationship between two (or more) measure results (e.g., the direction and strength of the relationship(s) between the measures). Typically a correlation analysis can be used to explore the hypothesized relationship, although other analytic techniques (e.g., regression) also could be used.
  - The developer reports :
    - For the hospital-level measure: "The performance measure score validity for the hospital performance measure score revealed that the Cronbach's alpha was .87.
    - For the provider-level measure: "The performance measure score predictive validity for the provider performance measure score revealed a regression line with an equation of y = 1.017x + 1.4964 and an R-squared value of 0.9802. "
  - The results reported for the provider-level analysis describes the predicted c-section rate based on 2004 data to the actual c-section rate from 2005-2007 data. This analysis, however, seems to address the predictive ability of the risk model rather than the validity of the measure score.

Questions for the Committee:

- *Were all critical data elements evaluated?*
- *Is the test sample adequate to generalize for widespread implementation?*
- *What hypotheses about relationships between this measure and others did the developer analyze for the score-level validity testing? Are these reasonable hypotheses?*
- *Do the results demonstrate sufficient validity so that conclusions about quality can be made?*
- *Do you agree that the score from this measure as specified is an indicator of quality?*

2b3-2b7. Threats to Validity

2b3. Exclusions:

The following exclusions are specified:

1. Gestational age at birth of less than 36 weeks 4 days or greater than 42 weeks and 3 days.
2. History of a prior cesarean birth.
3. Multiple gestation.
4. Not in vertex presentation.
5. Did not attempt to have a vaginal birth by attempting labor."

Questions for the Committee:

- *Are the exclusions consistent with the evidence?*
- *Are any patients or patient groups inappropriately excluded from the measure?*
- *Are the exclusions/exceptions of sufficient frequency and variation across providers to be needed (and outweigh the data collection burden)?*

2b4. Risk adjustment: **Risk-adjustment method** ☐ **None** ☐ **Statistical model** ☐ **Stratification**

☒ **Other: Cohort comparison**

Conceptual rationale for SDS factors included ? ☐ **Yes** ☒ **No**

SDS factors included in risk model? ☐ **Yes** ☒ **No**

Risk adjustment summary

- The developer describes the risk adjustment as " A cohort of 100 women with similar factors provides the inherent risk (expected cesarean birth rate) for each woman who gives birth." "A computer program was developed under the premise that women who labor with similar physical characteristics will have a similar inherent risk for cesarean birth. Averaging the inherent risks of a given group of women provides the average inherent risk for the group. The average inherent risk of the group creates the expected cesarean birth rate for that group and a comparison of the actual cesarean birth rate to the expected cesarean birth rate creates the risk adjustment."
- Risk factors: parity (nulliparous or multiparous), newborn weight, prepregnancy maternal body mass index, maternal age, maternal height, gestational age, maternal weight gain, and the type of onset of labor (spontaneous or induced). These factors were selected because they are objective, measured in every pregnancy and have been previously proven to significantly affect the risk that labor will result in a cesarean birth. Odds ratio tables are presented.
- Sociodemographic variables were not collected or analyzed.
- Cohort comparison development: The cohort for developing the "inherent risk" for several groups was 72,777 births in New York state in 2004. A total of 211,379 test records from 2007-2009 were assigned an inherent risk and each test record's obstetrical care provider's risk adjustment was calculated by comparing the obstetrical care provider's parity specific and onset of labor specific actual cesarean birth rate to their expected cesarean birth rate of their previous 100 birth records where the expected cesarean birth rate was determined by the average inherent risk of those records. After assigning the inherent risk and calculating the obstetrical care provider's risk adjustment these two values were used to predict the risk of cesarean birth for that test record.
- Risk model discrimination and calibration are presented. The R-squared value is 0.9802. R-squared is a statistical measure of how close the data are to the fitted regression line with values are between 0 and 100%: 0.98 means that the model explains 98% of the variability of the data around its mean. The Hosmer-Lemeshow test statistic is 348.34 (df=8; p-value=0.00). The Hosmer–Lemeshow test is a statistical test for goodness of fit for logistic regression models frequently in risk prediction models. The p-value of the H-L test indicates that the observed C-sections are not significantly different from those predicted by the model and that the overall fit is good.

Questions for the Committee:

- *Is an appropriate risk-adjustment strategy included in the measure?*
- *Was the development cohort from New York in 2004 sufficiently representative to evaluate inherent risk in current patients throughout the nation?*
- *Are the candidate and final variables included in the risk adjustment adequately described for the measure to be implemented?*
- *Are all of the risk adjustment variables present at the start of care?*

○ *Do you agree with the developer's rationale that there is no conceptual basis for adjusting this measure for SDS factors?*

2b5. Meaningful difference *(can statistically significant and clinically/practically meaningful differences in performance measure scores can be identified)*:

- The measure is specified for individual clinicians and hospitals.
- The developer instructs "A simple z-test is used to create a 95% confidence interval to determine if the selected cesarean birth measure is statistically significantly different than that of another entity or that of the average in the facility, county, state or nation."
- The developer states "Healthcare administrators can compare results at the level of the obstetrical care provider/group, facility, county, state or across the nation. A simple z-test is used to create a 95% confidence interval to determine if the selected cesarean birth measure is statistically significantly different than that of another entity or that of the average in the facility, county, state or nation."

The developer provided the following data from New York 2005-2007 (211,379 birth records, 2542 providers, 105 facilities) comparing to the average result for the state:

	#providers (%)	*% of births*	*#facilities (%)*	*% of births*
No different than state average	**1802 (70.9%)**	***55%***	***42 (40%)***	***28.6%***
Better than state average	***398 (15.6%)***	***20%***	***30 (29%)***	***38.0%***
Worse than state average	***342 (13.5%)***	***19%***	***33 (31%)***	***33.4%***

Question for the Committee:

○ *Does this measure identify meaningful differences about quality for clinicians and hospitals?*

2b6. Comparability of data sources/methods: NA

2b7. Missing Data

- Of 224,142 records used to assess performance, there were 12,763 (5.69%) that were missing one or more of the critical data elements or had unusual data. The cesarean birth rate of the deleted records was 14.1% which was not statistically significantly different from the 14.4% cesarean birth rate for the 211,379 birth records used to assess performance.

Guidance from the algorithm:
Specifications consistent with evidence (Box 1) →potential threats to validity assessed (Box 2) → empiric validity testing- uncertain conclusions (Box 3) → Testing of data elements (Box 10) →appropriate method for most, but not all, of the critical data elements (Box 11) →highest rating possible is moderate
Preliminary rating for validity: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 2: Scientific Acceptability of Measure Properties (including all 2a, 2b, and 2d)

2a1. & 2b1. Specifications

Comments:

Many items are known to be unreliable in BC data, please see details above.

**Data elements are unclear in this section which states 36+4 - 42+3 weeks gestation for the denominator.

Codes that should be considered for exclusion are for those conditions where a vaginal birth are contraindicated such as

HIV+, previa, and hx of myomectomy. If the data source is from the birth certificate data, in many cases this is inaccurate.**
Prediction seems much higher than anyone else's work, unclear why.
** Codes that should be considered for exclusion are for those conditions where a vaginal birth are contraindicated such as HIV+, previa, and hx of myomectomy.**

2a2. Reliability Testing
Comments:
No.
Data may be too old (2005-2007) since new evidence regarding labor management from the NICHD was released. Data from the birth certificate will be consistently unreliable.

2b2. Validity Testing
Comments:
Unclear.
Data from the birth certificate data will be consistently invalid. Gestational age selection is inconsistently written in the document.

2b3. Exclusions Analysis
2b4. Risk Adjustment/Stratification for Outcome or Resource Use Measures
2b5. Identification of Statistically Significant & Meaningful Differences In Performance
2b6. Comparability of Performance Scores When More Than One Set of Specifications
2b7. Missing Data Analysis and Minimizing Bias
Comments:
Poor data quality, differences is missing data, unclear that the outcome (CS) is truly a marker of poor labor management.
Should not be statistically different.
There are standard accepted methods to do O/E ratios. This measure does not seem to fit standard and well described methods. Unclear if this is a communication problem or a construct issue.

Criterion 3. Feasibility

3. Feasibility is the extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

This is a measure that requires a fee to calculate the measure results. NQF will consider such proprietary measures if there is full disclosure of all fees required. The Committee should factor in the costs in the evaluation of feasibility.

The developer has attached a fee schedule:

> *The pricing structure is dependent on the method of data entry.*
> *Data entry can be accomplished by sending birth data that has been transcribed onto an individual worksheet for each birth, in a digital format or via bridging software directly from the electronic health record.*
> *The cost per birth when birth data is transcribed onto a data worksheet and the worksheet is provided to Birthrisk.com, LLC is $12.00 per birth.*
> *The cost per birth when birth data is digitally provided to Birthrisk.com, LLC is $8.00 per birth.*
> *The cost per birth when a software bridge is created to provide data directly is $4.00 per birth plus the cost of establishing and maintaining the software bridge.*
> *These costs are required in order to provide data entry, maintenance of the dynamic database, auditing and reconciliation of data submitted along with providing reports.*

The developer notes that " The goal for timing and frequency of data collection is to have data submitted monthly so that performance measures can be made available within three months of when the births occurred."

Questions for the Committee:

- *Are the required data elements routinely generated and used during care delivery?*
- *Are the required data elements available in electronic form, e.g., EHR or other electronic sources?*
- *Is the data collection strategy ready to be put into operational use?*
- *What would be the typical costs to a clinician group or hospital to use this measure?*

Preliminary rating for feasibility: ☐ **High** ☐ **Moderate** ☒ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 3: Feasibility

3a. Byproduct of Care Processes
3b. Electronic Sources
3c. Data Collection Strategy

Comments:

Data is available but unreliable for variables used.
Birth certificate data is not reliable or valid in many states. Facilities may not be willing to pay for this measure to be calculated.

Criterion 4: Usability and Use

4. Usability and Use evaluate the extent to which audiences (e.g., consumers, purchasers, providers, policymakers) use or could use performance results for both accountability and performance improvement activities.

Current uses of the measure

Publicly reported? ☐ **Yes** ☒ **No**

Current use in an accountability program? ☐ **Yes** ☒ **No**

OR

Planned use in an accountability program? ☒ **Yes** ☐ **No**

The developer states that "This performance measure can be made public through the reporting mechanisms of the entities themselves or directly through the internet to the general public if so desired. Access to the general public could be through any device that can access the internet including mobile devices like smart phones."

Accountability program details new measure/no information
Improvement results new measure/no information
Unexpected findings (positive or negative) during implementation new measure/no information
Potential harms new measure/no information
Feedback: new measure/no information

Questions for the Committee:

- *How useable is this masure for clincians and hsopitals?*
- *How can the performance results be used to further the goal of high-quality, efficient healthcare?*
- *Do the benefits of the measure outweigh any potential unintended consequences?*

Preliminary rating for usability and use: ☐ **High** ☐ **Moderate** ☒ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 4: Usability and Use

4a. Accountability and Transparency
4b. Improvement
4c. Unintended Consequences

Comments:

May have unintended consequence of discouraging needed cesareans.
Not currently publicly reported.

Criterion 5: Related and Competing Measures

Related or competing measures

- 0471: PC-02 Cesarean Birth

Harmonization

The developer indicates that no attempt for harmonization has been made because he believes measure 0469 is flawed. His arguments are detailed here.

Pre-meeting public and member comments

-

NATIONAL QUALITY FORUM—Evidence (subcriterion 1a)

Measure Number (*if previously endorsed*)**:** Click here to enter NQF number
Measure Title: Birthrisk Cesarean Birth Measure
IF the measure is a component in a composite performance measure, provide the title of the Composite Measure here: Click here to enter composite measure #/ title

Date of Submission: 2/12/2016

Instructions

- *For composite performance measures:*
  - *A separate evidence form is required for each component measure unless several components were studied together.*
  - *If a component measure is submitted as an individual performance measure, attach the evidence form to the individual measure submission.*
- Respond to all questions as instructed with answers immediately following the question. All information needed to demonstrate meeting the evidence subcriterion (1a) must be in this form. An appendix of *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 10 pages (*incudes questions/instructions*; minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.

Note: The information provided in this form is intended to aid the Steering Committee and other stakeholders in understanding to what degree the evidence for this measure meets NQF's evaluation criteria.

1a. Evidence to Support the Measure Focus
The measure focus is evidence-based, demonstrated as follows:

- Health outcome: [3] a rationale supports the relationship of the health outcome to processes or structures of care. Applies to patient-reported outcomes (PRO), including health-related quality of life/functional status, symptom/symptom burden, experience with care, health-related behavior.
- Intermediate clinical outcome: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured intermediate clinical outcome leads to a desired health outcome.
- Process: [5] a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured process leads to a desired health outcome.
- Structure: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured structure leads to a desired health outcome.
- Efficiency: [6] evidence not required for the resource use component.

Notes

3. Generally, rare event outcomes do not provide adequate information for improvement or discrimination; however, serious reportable events that are compared to zero are appropriate outcomes for public reporting and quality improvement.
4. The preferred systems for grading the evidence are the U.S. Preventive Services Task Force (USPSTF) grading definitions and methods, or Grading of Recommendations, Assessment, Development and Evaluation (GRADE) guidelines.
5. Clinical care processes typically include multiple steps: assess → identify problem/potential problem → choose/plan intervention (with patient input) → provide intervention → evaluate impact on health status. If the measure focus is one step in such a multistep process, the step with the strongest evidence for the link to the desired outcome should be selected as the focus of measurement. Note: A measure focused only on collecting PROM data is not a PRO-PM.
6. Measures of efficiency combine the concepts of resource use and quality (see NQF's Measurement Framework: Evaluating Efficiency Across Episodes of Care; AQA Principles of Efficiency Measures).

1a.1.This is a measure of: (*should be consistent with type of measure entered in De.1*)
Outcome

☒ Health outcome: Cesarean Birth

☐Patient-reported outcome (PRO): Click here to name the PRO

PROs include HRQoL/functional status, symptom/symptom burden, experience with care, health-related behaviors

☐ Intermediate clinical outcome (*e.g., lab value*): Click here to name the intermediate outcome

☐ Process: Click here to name the process

☐ Structure: Click here to name the structure

☐ Other: Click here to name what is being measured

HEALTH OUTCOME/PRO PERFORMANCE MEASURE *If not a health outcome or PRO, skip to 1a.3*

1a.2. Briefly state or diagram the path between the health outcome (or PRO) and the healthcare structures, processes, interventions, or services that influence it.

The measure focus is the effect of the obstetrical care provider´s labor management strategies on a woman´s risk that labor will result in a cesarean birth. A woman's risk that labor will result in a cesarean birth is significantly affected by her unique combination of physical characteristics and the size of her baby. Obstetrical care providers have attempted to assist women in labor by introducing labor management strategies. Measuring the effect of the obstetrical care provider's labor management strategies on a woman's inherent risk for cesarean birth will identify those strategies that can be used to decrease the number of cesarean births. A decrease in the number of cesarean births will decrease maternal morbidity and the cost of maternity care.

1a.2.1. State the rationale supporting the relationship between the health outcome (or PRO) to at least one healthcare structure, process, intervention, or service (*i.e., influence on outcome/PRO*).

There are many different labor management strategies that have been used over the years to assist women who are in labor. Some of these strategies can decrease a woman's inherent risk that labor will result in a cesarean birth and others can increase the risk. For example, the use of forceps can decrease a woman's inherent risk for a cesarean birth whereas an impatient obstetrical care provider may increase a woman's inherent risk.

Note: For health outcome/PRO performance measures, no further information is required; however, you may provide evidence for any of the structures, processes, interventions, or service identified above.

INTERMEDIATE OUTCOME, PROCESS, OR STRUCTURE PERFORMANCE MEASURE

1a.3. Briefly state or diagram the path between structure, process, intermediate outcome, and health outcomes. Include all the steps between the measure focus and the health outcome.

1a.3.1. What is the source of the systematic review of the body of evidence that supports the performance measure?

☐ Clinical Practice Guideline recommendation – ***complete sections 1a.4, and 1a.7***

☐ US Preventive Services Task Force Recommendation – ***complete sections 1a.5 and 1a.7***

☐ Other systematic review and grading of the body of evidence (*e.g., Cochrane Collaboration, AHRQ Evidence Practice Center*) – ***complete sections 1a.6 and 1a.7***

☒ Other – ***complete section 1a.8***

Please complete the sections indicated above for the source of evidence. You may skip the sections that do not apply.

1a.4. CLINICAL PRACTICE GUIDELINE RECOMMENDATION

1a.4.1. Guideline citation (*including date*) and **URL for guideline** (*if available online*):

1a.4.2. Identify guideline recommendation number and/or page number and **quote verbatim, the specific guideline recommendation**.

1a.4.3. Grade assigned to the quoted recommendation with definition of the grade:

1a.4.4. Provide all other grades and associated definitions for recommendations in the grading system. (*Note: If separate grades for the strength of the evidence, report them in section 1a.7.*)

1a.4.5. Citation and URL for methodology for grading recommendations (*if different from 1a.4.1*)**:**

1a.4.6. If guideline is evidence-based (rather than expert opinion), are the details of the quantity, quality, and consistency of the body of evidence available (e.g., evidence tables)?

☐ Yes → ***complete section 1a.7***

☐ No → ***report on another systematic review of the evidence in sections 1a.6 and 1a.7; if another review does not exist, provide what is known from the guideline review of evidence in 1a.7***

1a.5. UNITED STATES PREVENTIVE SERVICES TASK FORCE RECOMMENDATION

1a.5.1. Recommendation citation (*including date*) and **URL for recommendation** (*if available online*):

1a.5.2. Identify recommendation number and/or page number and **quote verbatim, the specific recommendation**.

1a.5.3. Grade assigned to the quoted recommendation with definition of the grade:

1a.5.4. Provide all other grades and associated definitions for recommendations in the grading system. (*Note: the grading system for the evidence should be reported in section 1a.7.*)

1a.5.5. Citation and URL for methodology for grading recommendations (*if different from 1a.5.1*)**:**

Complete section 1a.7

1a.6. OTHER SYSTEMATIC REVIEW OF THE BODY OF EVIDENCE

1a.6.1. Citation (*including date*) and **URL** (*if available online*):

1a.6.2. Citation and URL for methodology for evidence review and grading (*if different from 1a.6.1*)**:**

Complete section 1a.7

1a.7. FINDINGS FROM SYSTEMATIC REVIEW OF BODY OF THE EVIDENCE SUPPORTING THE MEASURE

If more than one systematic review of the evidence is identified above, you may choose to summarize the one (or more) for which the best information is available to provide a summary of the quantity, quality, and consistency of the body of evidence. Be sure to identify which review is the basis of the responses in this section and if more than one, provide a separate response for each review.

1a.7.1. What was the specific structure, treatment, intervention, service, or intermediate outcome addressed in the evidence review?

1a.7.2. Grade assigned for the quality of the quoted evidence with definition of the grade:

1a.7.3. Provide all other grades and associated definitions for strength of the evidence in the grading system.

1a.7.4. What is the time period covered by the body of evidence? (*provide the date range, e.g., 1990-2010*). Date range: Click here to enter date range

QUANTITY AND QUALITY OF BODY OF EVIDENCE

1a.7.5. How many and what type of study designs are included in the body of evidence? (*e.g., 3 randomized controlled trials and 1 observational study*)

1a.7.6. What is the overall quality of evidence across studies in the body of evidence? (*discuss the certainty or confidence in the estimates of effect particularly in relation to study factors such as design flaws, imprecision due to small numbers, indirectness of studies to the measure focus or target population*)

ESTIMATES OF BENEFIT AND CONSISTENCY ACROSS STUDIES IN BODY OF EVIDENCE

1a.7.7. What are the estimates of benefit—magnitude and direction of effect on outcome(s) across studies in the body of evidence? (*e.g., ranges of percentages or odds ratios for improvement/ decline across studies, results of meta-analysis, and statistical significance*)

1a.7.8. What harms were studied and how do they affect the net benefit (benefits over harms)?

UPDATE TO THE SYSTEMATIC REVIEW(S) OF THE BODY OF EVIDENCE

1a.7.9. If new studies have been conducted since the systematic review of the body of evidence, provide for each new study: 1) citation, 2) description, 3) results, 4) impact on conclusions of systematic review.

1a.8 OTHER SOURCE OF EVIDENCE

If source of evidence is NOT from a clinical practice guideline, USPSTF, or systematic review, please describe the evidence on which you are basing the performance measure.

In 1964 Dr. Edward H. Bishop published his landmark work titled "Pelvic Scoring for Elective Induction" as a practice guideline for the labor management strategy of elective induction of labor [1]. Countless studies and numerous guidelines regarding the results of labor management strategies have since been developed [2-20]. In 2000 the American College of Obstetrician and Gynecologist's task force on cesarean birth published an Evaluation of Cesarean Delivery which had 256 references [2]. In 2014 the American College of Obstetricians and Gynecologists and the Society for Maternal-Fetal Medicine published a consensus on the safe prevention of the primary cesarean delivery with 114 references [20]. These studies provide a tremendous body of evidence that not only do the labor management strategies of the obstetrical care provider significantly affect a woman's risk for cesarean birth but that each woman's risk that labor will result in a cesarean birth is significantly affected by her physical characteristics and the size of her baby. Maternal physical characteristics like parity, prepregnancy body mass index, age, height, gestational age, pregnancy weight gain as well as fetal birth weight and induction of labor have all been confirmed in numerous studies to significantly affect a woman's risk that labor will result in a cesarean birth. All of these factors must be taken into

account in order to create an accurate measure of the effect of the obstetrical care provider's labor management strategies.

Even though the body of evidence consistently confirms that these factors significantly affect a woman's risk that labor will result in a cesarean birth there is a study by Caceres et al. that has been recently referenced as providing evidence that a woman's physical characteristics and the size of her baby can be ignored. The study states "One implication of this finding is that presenting hospital-specific cesarean rates for NTSV births might be appropriate without further case-mix adjustment" [21]. Critical review of the work by Caceres et al. reveals poorly selected risk factors, questionable quality of data collection, the work is not similar to the studies that were referenced and there is no attempt to explain the obvious contradiction between proving that risk factors significantly affect the risk of cesarean birth and then implying that they can be ignored when comparing outcome. In fact, one of the authors (Declercq E) of the Caceres study recently published the results of the effect of prepregnancy maternal body mass index on over two million women without a history of a prior cesarean birth who gave birth in 2012 to a single term (37-41 weeks) vertex baby [22]. The recent study by Declercq et al. concluded that "… prepregnancy obesity was found to represent an independent risk factor for primary cesareans, even after controlling for maternal demographic characteristics and medical risk factors." The result of this larger more recent study contradicts the implication made by Caceres et al.

Data from millions of women who have given birth has recently become available [23]. Analysis of the data from over two million women who attempted labor in 2011 reveals that inducing labor in a five foot two inch 36 year old nulliparous woman with a starting weight of 175 lbs. who has gained 42 lbs. and is carrying a 4,000 gram baby has a cesarean birth rate of approximately 70%. Whereas a five foot four inch 18 year old nulliparous woman with a starting weight of 115 lbs. who has gained 30 lbs. carrying a 3,500 gram baby who arrives in spontaneous labor has a cesarean birth rate of approximately 7%. This tenfold difference in the rate of cesarean birth due to the physical characteristics of the mother and the size of her baby reveals that any attempt to use an unadjusted NTSV cesarean birth rate as a measure of the effect of the obstetrical care provider's labor management strategies will be fatally flawed.

The overwhelming majority of the studies present in the body of evidence consistently confirm that the physical characteristics of the mother, the size of her baby and the labor management strategies used by the obstetrical care provider can all significantly affect a woman's risk that labor will result in a cesarean birth. Discovering the labor management strategies that result in the appropriate cesarean birth rate for a population of women has a tremendous benefit for the women who are giving birth as well as to the overall cost of maternity care.

1a.8.1 What process was used to identify the evidence?

The developer of this measure is Gustavo San Román, MD, who is a board certified obstetrician gynecologist that has been managing laboring patients since 1986. In 2006, Dr. San Román began to collect data and review the literature in an attempt to create a better cesarean birth measure. Since 2006, a continuous review of the literature has been conducted with representative articles being listed below.

1a.8.2. Provide the citation and summary for each piece of evidence.

Each of the following references provide evidence that both the physical characteristics of the mother and her baby as well as the effect of the obstetrical care provider affect a woman's risk for cesarean birth.

1. Bishop EH, Pelvic Scoring for Elective Induction. Obstet Gynecol 1964;24:166-8-7
2. American College of Obstetricians and Gynecologists Task Force on Cesarean Delivery. Evaluation of cesarean delivery. Washington (DC): American College of Obstetricians and Gynecologists; 2000. p. 1-59
3. Kozhimannil KB, Law MR, Virnig BA. Cesarean Delivery Rates Vary 10-Fold Among US Hospitals; Reducing Variation May Address Quality, Cost Issues. Health affairs (Project Hope). 2013;32(3):527-535
4. Bailit JL, Love TE, Mercer B. Rising cesarean rates: are patients sicker? Am J Obstet Gynecol 2004;191:800-3.

5. Chen G, Uryasev S, Young T. On prediction of the cesarean delivery risk in a large private practice. Am J Obstet Gynecol 2004;191:617-25
6. Peregrine E, O'Brien P, Omar R, Jauniaux E. Clinical and Ultrasound Parameters to Predict the Risk of Cesarean Delivery After Induction of Labor. Obstet Gynecol 2006;107:227-33
7. Wilkes P, Wolf D, Kronbach D, Kunze M, Gibbs R. Risk Factors for Cesarean Delivery at Presentation of Nulliparous Patients in Labor. Obstet Gynecol 2003;102:1352-7
8. Bergholt T, Lim L, Jorgensen J, Robson M. Maternal body mass index in the first trimester and risk of cesarean delivery in nulliparous women in spontaneous labor. Am J Obstet Gynecol 2007;196(2):163.e1-5
9. Maslow A, Sweeny A. Elective Induction of Labor as a Risk Factor for Cesarean Delivery Among Low-Risk Women at Term. Obstet & Gynecol 2000;95:917-22
10. National Collaborating Centre for Women's and Children's Health. Induction of Labour – Clinical Guideline. 2nd ed. Regent's Park, London: RCOG Press; 2008
11. Ehrenthal DB, Jiang X, Strobino DM. Labor Induction and the Risk of a Cesarean Delivery Among Nulliparous Women at Term. Obstet Gynecol 2010;116:35-42
12. Luthy DA, Malmgren JA, Zingheim RW, Leninger CJ. Physician contribution to a cesarean delivery risk model. Am J Obstet Gynecol 2003;188:1579-87.
13. Coonrod DV, Drachman D, Hobson P, et al. Nulliparous term singleton vertex cesarean delivery rates: institutional and individual level predictors. Am J Obstet Gynecol 2008;198:694.e1-694.e11
14. Barber EL, Lundsberg LS, Belanger K, Pettker CM, Funai EF, Illuzzi JL. Indications Contributing to the Increasing Cesarean Delivery Rate. Obstet Gynecol 2011;118:29-38
15. Main E, Moore D, Farrel B, Schimmel L, Altman R, Abrahams C, et al. Is there a useful cesarean birth measure? Assessment of the nulliparous term singleton vertex cesarean birth rate as a tool for obstetric quality improvement. Am J Obstet Gynecol 2006;194:1644-52
16. Robson MS. Classification of caesarean sections. Fetal and Maternal Medicine Review. 2001;12:23–39.
17. Goyert GL, Bottoms SF, Treadwell MC, Nehra PC. The Physician Factor in Cesarean Birth Rates. N. Engl J Med 1989; 320:706-9
18. Clark SL, Belfort MA, Hankins GDV, Meyers JA, Houser FM (2007) Variation in the rates of operative delivery in the United States. Am J Obstet Gynecol 196:526.e1–5
19. Brennan DJ, Robson MS, Murphy M, et al. Comparative analysis of international cesarean delivery rates using 10-group classification identifies significant variation in spontaneous labor. Am J Obstet Gynecol 2009;201:308.e1-8.
20. Caughey AB, Cahill AG, Guise JM, Rouse DJ. Safe Prevention of the Primary Cesarean Delivery. Obstet Gynecol 2014;123:693-711
21. Caceres IA, Arcaya M, Declercq E, Belanoff CM, Janakiraman V, et al. (2013) Hospital Differences in Cesarean Deliveries in Massachusetts (US) 2004–2006: The Case against Case-Mix Artifact. PLoS ONE 8(3): e57817
22. Declercq E, MacDorman M, Osterman M, Belanoff C, Iverson R. Prepregnancy Obesity and Primary Cesareans among Otherwise Low-Risk Mothers in 38 U.S. States in 2012. Birth 2015;42(4):309-18
23. Centers for Disease Control and Prevention. 2011 Birth Data Files. Available at http://www.cdc.gov/nchs/data_access/Vitalstatsonline.htm. Retrieved

1. Evidence, Performance Gap, Priority – Importance to Measure and Report

Extent to which the specific measure focus is evidence-based, important to making significant gains in healthcare quality, and improving health outcomes for a specific high-priority (high-impact) aspect of healthcare where there is variation in or overall less-than-optimal performance. ***Measures must be judged to meet all subcriteria to pass this criterion and be evaluated against the remaining criteria.***

1a. Evidence to Support the Measure Focus – See attached Evidence Submission Form
Birthrisk_Cesarean_Birth_Measure-evidence_submission_form.docx

1b. Performance Gap

Demonstration of quality problems and opportunity for improvement, i.e., data demonstrating:

- considerable variation, or overall less-than-optimal performance, in the quality of care across providers; and/or
- disparities in care across population groups.

1b.1. **Briefly explain the rationale for this measure** *(e.g., the benefits or improvements in quality envisioned by use of this measure)*
Healthcare administrators, hospital administrators and obstetrical care providers do not currently have an accurate method for determining if their labor management strategies are better or worse than the average. The reason that an accurate method is not available is because the physical process of labor makes it impossible to assess the effect of the obstetrical care provider´s labor management strategies unless the unique combination of physical characteristics of the mother and the size of her baby are taken into account.

Empowering obstetrical care providers with an accurate cesarean birth measure will allow them to analyze which labor management strategies work best at reaching the appropriate cesarean birth rate for their population of patients. Hospital administrators will have the ability to conduct and monitor quality improvement projects aimed at achieving the appropriate cesarean birth rate for their institution by analyzing the results achieved by each obstetrical care provider in their hospital. Healthcare administrators will be able to monitor the various organizational units under their jurisdiction in order to identify centers of excellence as well as centers that need improvement. With the ability to compare organizational units to the average in the state or nation, healthcare administrators could initiate a state/nation wide incentive/disincentive program for entities that are statistically significantly better/worse than the average in the state/nation.

1b.2. **Provide performance scores on the measure as specified (current and over time) at the specified level of analysis.** *(This is required for endorsement maintenance. Include mean, std dev, min, max, interquartile range, scores by decile. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included). This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*
Evaluation of the performance score was evaluated at the facility level for the time period of 2005 to 2007. There were 105 facilities assigned a cesarean birth measure however two of these facilities had less than 10 births during the time period and were not included in the analysis. The mean facility performance score was 15.57% with a minimum score of 7% and a maximum of 29%. The standard deviation was 4.52 and the interquartile range was 5% (Q1=13%, Q3=18%). There were 211,374 births in the analysis with a mean number of births per facility of 2,052 and a range of 64 to 9,364.

Evaluation of the performance score was evaluated at the provider level for the time period of 2005 to 2007. There were 2,542 providers assigned a cesarean birth measure however 650 of these providers had less than 10 births during the time period and were not included in the analysis. The mean provider performance score was 14.31% with a minimum score of 0% and a maximum of 71%. The standard deviation was 11.07 and the interquartile range was 17% (Q1=4%, Q3=21%). There were 209,300 births in the analysis with a mean number of births per provider of 111 and a range of 10 to 1,055.

Evaluation of performance scores were evaluated at the facility with the most births (9,364) for the time period of 2005 to 2007. This facility had a performance score of 17.01%. There were 44 births to providers who had less than 10 births during the time period and were not included in the analysis. The mean provider performance score was 17.11% with a minimum score of 0% and a maximum of 33%. The nine providers who had performance measures that were significantly worse than average were analyzed separately from the remaining 59 providers. The mean provider performance score for the nine providers was 23.62% and these providers attended 1,699 (18%) of the births. The mean provider performance score for the 59 remaining providers was 16.22% and these 59 providers attended 7,621 (82%) of the births.

1b.3. **If no or limited performance data on the measure as specified is reported in 1b2, then provide a summary of data from the**

literature that indicates opportunity for improvement or overall less than optimal performance on the specific focus of measurement.

As far back as 1989 Goyert et al. concluded that individual practice style may be an important determinant of the wide variations in the rates of cesarean delivery among obstetricians [1]. In June of 2007 Clark et al. reported on 220,000 births and concluded that rates of operative delivery in the United States are highly variable and suggest a pattern of almost random decision making [2]. In March 2009 Kozhimannil et al. found that cesarean rates varied tenfold across 593 U.S. hospitals [3]. The significant variation in cesarean birth rates is illustrated every year in the National Vital Statistics Reports [4].

The variation in cesarean birth rates is not a problem unique to the United States. In 2009 Brennan DJ et al. concluded that analysis of international obstetric cesarean practice identifies wide variations in women in spontaneous cephalic term labor [5]. However, to date there are no published studies that take into account the unique combination of the physical characteristics of the mother and the size of her baby. Unfortunately, if a study does not take into account the unique combination of the physical characteristics of the mother and the size of her baby then the reported performance gap may not be accurate. Only a measure that can accurately account for the unique combination of a woman's physical characteristics and the size of her baby will be able to determine how much of the variation in rates is dependent on the physical characteristics of the women who are giving birth and how much of the variation in rates is due to the labor management strategies being used by the obstetrical care provider. The research that resulted in the development of the Birthrisk Cesarean Birth Measure shows that there is a significant performance gap with some obstetrical care providers performing 20 to 30 more cesarean births per 100 women in labor when compared to the average in their state [6].

1. Goyert GL, Bottoms SF, Treadwell MC, Nehra PC. The Physician Factor in Cesarean Birth Rates. N. Engl J Med 1989; 320:706-9
2. Clark SL, Belfort MA, Hankins GDV, Meyers JA, Houser FM (2007) Variation in the rates of operative delivery in the United States. Am J Obstet Gynecol 196:526.e1–5
3. Kozhimannil KB, Law MR, Virnig BA. Cesarean Delivery Rates Vary Tenfold Among US Hospitals; Reducing Variation May Address Quality and Cost Issues. Health Affairs, 32, no.3 (2013):527-35
4. Hamilton BE, Martin JA, Osterman MJK, Curtin SC. Births: Preliminary data for 2014. National vital statistics reports; vol 64 no 6. Hyattsville, MD: National Center for Health Statistics. Released June 17, 2015
5. Brennan DJ, Robson MS, Murphy M, et al. Comparative analysis of international cesarean delivery rates using 10-group classification identifies significant variation in spontaneous labor. Am J Obstet Gynecol 2009;201:308.e1-8.
6. Birthrisk.com, LLC., Information for Healthcare Professionals. Available at https://www.birthrisk.com/Public/BirthriskVideos1.aspx. Retrieved January 16, 2016

1b.4. Provide disparities data from the measure as specified (current and over time) by population group, e.g., by race/ethnicity, gender, age, insurance status, socioeconomic status, and/or disability. *(This is required for endorsement maintenance. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities include.) This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*

N/A

1b.5. If no or limited data on disparities from the measure as specified is reported in 1b4, then provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement. Include citations.

Ehrenthal et al. conducted a retrospective cohort study of cesarean births among nulliparous women delivering a live, singleton, vertex pregnancy at term [1]. They used clinical data from electronic hospital obstetric records at a large, regional, obstetric hospital, approximating a population-based cohort. Their results showed a greater odds of cesarean birth from the following sociodemographic characteristics: black race, marital status, patient type (private or service), insurance type (Medicaid or private), and age older than 35 years.

However, Declercq et al. used a sample of 2,233,144 women who had a singleton, vertex, term (37–41 weeks) birth in 2012 and no prior cesarean to demonstrate that prepregnancy obesity was found to represent an independent risk factor for primary cesareans [2]. Their results showed that obesity rates were highest among American Indian and Alaska Native (32.5%) and non-Hispanic black mothers (30.5%).

Unfortunately, there are no studies that address disparities in care due to sociodemographic factors while adequately adjusting for the unique combination of a woman's physical characteristics and the size of her baby. For example, in the study by Ehrenthal et al. the continuous variables of fetal weight, prepregnancy body mass index, maternal age, gestational age and maternal weight gain are represented as ordinal variables. Representing a continuous variable as an ordinal variable could decrease the significance of that variable in the multiple logistic regression analysis. Ehrenthal's finding that black race had greater odds of having a cesarean birth was based on their multiple logistic regression analysis.

Identifying true disparities in care associated with sociodemographic factors will require the use of a method that takes into account the unique combination of a woman's physical characteristics and the size of her baby. The Birthrisk Cesarean Birth Measure provides that method.

1. Ehrenthal DB, Jiang X, Strobino DM. Labor Induction and the Risk of a Cesarean Delivery Among Nulliparous Women at Term. Obstet Gynecol 2010;116:35-42
2. Declercq E, MacDorman M, Osterman M, Belanoff C, Iverson R. Prepregnancy Obesity and Primary Cesareans among Otherwise Low-Risk Mothers in 38 U.S. States in 2012. Birth 2015;42(4):309-18

1c. High Priority (previously referred to as High Impact)

The measure addresses:

- a specific national health goal/priority identified by DHHS or the National Priorities Partnership convened by NQF; OR
- a demonstrated high-priority (high-impact) aspect of healthcare (e.g., affects large numbers of patients and/or has a substantial impact for a smaller population; leading cause of morbidity/mortality; high resource use (current and/or future); severity of illness; and severity of patient/societal consequences of poor quality).

1c.1. Demonstrated high priority aspect of healthcare

Affects large numbers, High resource use, Patient/societal consequences of poor quality

1c.2. If Other:

1c.3. Provide epidemiologic or resource use data that demonstrates the measure addresses a high priority aspect of healthcare. List citations in 1c.4.

According to the National Center for Health Statistics (NCHS) cesarean birth is the most commonly performed inpatient surgical procedure in the United States [1]. The NCHS also reports that the overall cesarean birth rate in the United States is up nearly 60% (almost twelve percentage points) since 1996 with an all-time high of 32.9% of all births in 2009 [2]. Recent efforts to address the rising rate have only achieved a 2% decrease (seven tenths of one percentage point) from 2009 to 2014 [2-4]. Each percentage point in the United States cesarean birth rate represents approximately 40,000 women who undergo a surgical procedure which has increased morbidity and costs when compared to a vaginal birth [5-7]. A recent study concluded that National cesarean delivery rates of up to approximately 19 per 100 live births were associated with lower maternal or neonatal mortality which suggests that a National cesarean birth rate of approximately 19% would be appropriate [8].

The majority of pregnant women will attempt labor with the goal of achieving a vaginal birth. The pregnant women who do not attempt labor presumably schedule a cesarean birth as the result of a well-informed discussion with their obstetrical care provider. If all of the scheduled cesarean births are appropriate then in order to achieve the appropriate cesarean birth rate the focus must be on the women who labor. Understanding the factors that affect the cesarean birth rate for women who labor will first and foremost require the acknowledgement that labor is a physical process. It has been shown that a woman's physical characteristics and the size of her baby affect her risk that labor will result in a cesarean birth [9-17]. However, the risk that labor will result in a cesarean birth is also significantly affected by the obstetrical care provider's labor management strategies [18-20]. These strategies can either increase or decrease a woman's risk. For example, the use of forceps can decrease a woman's risk for a cesarean birth whereas an impatient obstetrical care provider may increase a woman's risk.

The goal of a cesarean birth measure is to measure the effect of the labor management strategies applied by obstetrical care providers to each woman's risk that labor will result in a cesarean birth, keeping in mind that the obstetrical care provider's labor management strategies can either increase or decrease a woman's risk. In order to be able to accurately measure the effect of labor management strategies it is mandatory to have an accurate risk adjustment method that accounts for each woman's unique combination of physical characteristics and the size of her baby. Prior attempts at creating cesarean birth measures have used established methods for risk adjustment such as direct standardization or grouping by risk factors [21,22]. Direct standardization is easy to use but will create meaningless results if any of the risk strata are empty or contain very few patients. This makes direct standardization impractical for comparing obstetrical care providers. Grouping women by risk factors that ignore the physical characteristics of the women who are giving birth will yield results that are dependent on the physical characteristics of the women in each group rather than a measure of the labor management strategies applied by the obstetrical care provider. It is for these reasons that no prior cesarean birth measure has ever been validated. Only an accurate cesarean birth measure will allow for the discovery of the labor management strategies that are best at reaching the appropriate cesarean birth rate for any given population of women.

1c.4. **Citations for data demonstrating high priority provided in 1a.3**

1. National Center for Health Statistics. Fast Stats - Inpatient Surgery. Available at http://www.cdc.gov/nchs/fastats/inpatient-surgery.htm. Retrieved January 7, 2016
2. Hamilton BE, Martin JA, Osterman MJK, Curtin SC. Births: Preliminary data for 2014. National vital statistics reports; vol 64 no 6. Hyattsville, MD: National Center for Health Statistics. Released June 17, 2015
3. The Joint Commission. Specifications Manual for Joint Commission National Quality Core Measures version 2012B. Available at https://manual.jointcommission.org/releases/TJC2015B2/MIF0167.html. Retrieved January 16, 2016
4. Department of Health and Human Services. Healthy People 2020 summary of objectives, maternal, infant, and child health. Available at http://www.healthypeople.gov/node/4900/data_detailsMaternalChildHealth.pdf. Retrieved January 16, 2016
5. American College of Obstetricians and Gynecologists Task Force on Cesarean Delivery. Evaluation of cesarean delivery. Washington (DC): American College of Obstetricians and Gynecologists; 2000. p. 1-59
6. Kozhimannil KB, Law MR, Virnig BA. Cesarean Delivery Rates Vary 10-Fold Among US Hospitals; Reducing Variation May Address Quality, Cost Issues. Health affairs (Project Hope). 2013;32(3):527-535
7. National Center for Health Statistics. Health, United States, 2013: with special feature on prescription drugs. Hyattsville, MD. 2014; http://www.cdc.gov/nchs/data/hus/hus13.pdf. Accessed January 6, 2016.
8. Molina G, et. al. Relationship Between Cesarean Delivery Rate and Maternal and Neonatal Mortality. JAMA 2015;314(21):2263-2270
9. Bailit JL, Love TE, Mercer B. Rising cesarean rates: are patients sicker? Am J Obstet Gynecol 2004;191:800-3.
10. Chen G, Uryasev S, Young T. On prediction of the cesarean delivery risk in a large private practice. Am J Obstet Gynecol 2004;191:617-25
11. Peregrine E, O'Brien P, Omar R, Jauniaux E. Clinical and Ultrasound Parameters to Predict the Risk of Cesarean Delivery After Induction of Labor. Obstet Gynecol 2006;107:227-33
12. Wilkes P, Wolf D, Kronbach D, Kunze M, Gibbs R. Risk Factors for Cesarean Delivery at Presentation of Nulliparous Patients in Labor. Obstet Gynecol 2003;102:1352-7
13. Bergholt T, Lim L, Jorgensen J, Robson M. Maternal body mass index in the first trimester and risk of cesarean delivery in nulliparous women in spontaneous labor. Am J Obstet Gynecol 2007;196(2):163.e1-5
14. Declercq E, MacDorman M, Osterman M, Belanoff C, Iverson R. Prepregnancy Obesity and Primary Cesareans among Otherwise Low-Risk Mothers in 38 U.S. States in 2012. Birth 2015;42(4):309-18
15. Maslow A, Sweeny A. Elective Induction of Labor as a Risk Factor for Cesarean Delivery Among Low-Risk Women at Term. Obstet & Gynecol 2000;95:917-22
16. National Collaborating Centre for Women's and Children's Health. Induction of Labour – Clinical Guideline. 2nd ed. Regent's Park, London: RCOG Press; 2008
17. Ehrenthal DB, Jiang X, Strobino DM. Labor Induction and the Risk of a Cesarean Delivery Among Nulliparous Women at Term. Obstet Gynecol 2010;116:35-42
18. Luthy DA, Malmgren JA, Zingheim RW, Leninger CJ. Physician contribution to a cesarean delivery risk model. Am J Obstet Gynecol 2003;188:1579-87.
19. Coonrod DV, Drachman D, Hobson P, et al. Nulliparous term singleton vertex cesarean delivery rates: institutional and individual level predictors. Am J Obstet Gynecol 2008;198:694.e1-694.e11
20. Barber EL, Lundsberg LS, Belanger K, Pettker CM, Funai EF, Illuzzi JL. Indications Contributing to the Increasing Cesarean Delivery Rate. Obstet Gynecol 2011;118:29-38
21. Main E, Moore D, Farrel B, Schimmel L, Altman R, Abrahams C, et al. Is there a useful cesarean birth measure? Assessment of the nulliparous term singleton vertex cesarean birth rate as a tool for obstetric quality improvement. Am J Obstet Gynecol 2006;194:1644-52
22. Robson MS. Classification of caesarean sections. Fetal and Maternal Medicine Review. 2001;12:23–39.

1c.5. **If a PRO-PM (e.g. HRQoL/functional status, symptom/burden, experience with care, health-related behaviors), provide evidence that the target population values the measured PRO and finds it meaningful.** *(Describe how and from whom their input was obtained.)*

N/A

2. Reliability and Validity—Scientific Acceptability of Measure Properties

Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. ***Measures must be judged to meet the subcriteria for both reliability and validity to pass this criterion and be evaluated against the remaining criteria.***

2a.1. Specifications The measure is well defined and precisely specified so it can be implemented consistently within and across organizations and allows for comparability. eMeasures should be specified in the Health Quality Measures Format (HQMF) and the Quality Data Model (QDM).

De.5. **Subject/Topic Area** *(check all the areas that apply):*
Perinatal and Reproductive Health : Perinatal

De.6. **Cross Cutting Areas** *(check all the areas that apply):*

S.1. **Measure-specific Web Page** *(Provide a URL link to a web page specific for this measure that contains current detailed specifications including code lists, risk model details, and supplemental materials. Do not enter a URL linking to a home page or to general information.)*
https://www.birthrisk.com/Public/NQF.aspx

S.2a. **If this is an eMeasure**, HQMF specifications must be attached. Attach the zipped output from the eMeasure authoring tool (MAT) - if the MAT was not used, contact staff. (Use the specification fields in this online form for the plain-language description of the specifications)
This is not an eMeasure **Attachment:**

S.2b. **Data Dictionary, Code Table, or Value Sets** *(and risk model codes and coefficients when applicable) must be attached. (Excel or csv file in the suggested format preferred - if not, contact staff)*
No data dictionary **Attachment:**

S.3. **For endorsement maintenance**, please briefly describe any changes to the measure specifications since last endorsement date and explain the reasons.
N/A

S.4. **Numerator Statement** *(Brief, narrative description of the measure focus or what is being measured about the target population, i.e., cases from the target population with the target process, condition, event, or outcome)*
IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
Number of cesarean births.

S.5. **Time Period for Data** *(What is the time period in which data will be aggregated for the measure, e.g., 12 mo, 3 years, look back to August for flu vaccination? Note if there are different time periods for the numerator and denominator.)*
The time period has a minimum value of one day. Public reporting and external bench-marking can be provided monthly, quarterly, annually or any other time period with a minimum value of one day.

S.6. **Numerator Details** *(All information required to identify and calculate the cases from the target population with the target process, condition, event, or outcome such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
IF an OUTCOME MEASURE, describe how the observed outcome is identified/counted. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
The number of births with Method of Delivery reported as Cesarean. U.S. Standard Certificate of Birth item number 46 (METHOD OF DELIVERY), processing variable: ROUT=4.

S.7. **Denominator Statement** *(Brief, narrative description of the target population being measured)*
Women without a history of a prior cesarean birth who attempted labor and gave birth to a single baby in vertex presentation between 37 and 42 weeks of gestation.

S.8. **Target Population Category** *(Check all the populations for which the measure is specified and tested if any):*

S.9. **Denominator Details** *(All information required to identify and calculate the target population/denominator such as definitions, specific data collection items/responses , code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
The denominator is all of the women who gave birth during the specified time period as determined by an existing Certificate of Birth. Data collection items from the U.S. Standard Certificate of Birth are listed by Item Number, Description and (Processing Variable(s)):

Item 2 TIME OF BIRTH (TB)
Item 4 DATE OF BIRTH - infant (IDOB_YR, IDOB_MO, IDOB_DY)
Item 5 FACILITY NAME (FNAME)
Item 6 CITY, TOWN OR LOCATION OF BIRTH (FLOC)
Item 7 COUNTY OF BIRTH (CNAME)
Item 8b DATE OF BIRTH - mother (MDOB_YR, MDOB_MO, MDOB_DY)
Item 27 ATTENDANT'S NAME, TITLE, AND NPI (ATTENDN, NPI)
Item 28 MOTHER TRANSFERRED FOR MATERNAL MEDICAL OR FETAL INDICATIONS FOR DELIVERY? (TRAN, NFACL)
Item 31 MOTHER'S HEIGHT (HFT, HIN)
Item 32 MOTHER'S PREPREGNANCY WEIGHT (PWGT)
Item 33 MOTHER'S WEIGHT AT DELIVERY (DWGT)
Item 35a NUMBER OF PREVIOUS LIVE BIRTHS - NOW LIVING (PLBL)
Item 35b NUMBER OF PREVIOUS LIVE BIRTHS - NOW DEAD (PLBD)
Items 41 RISK FACTORS IN THIS PREGNANCY - Mother had a previous cesarean delivery (PCES)
Item 44 ONSET OF LABOR - Precipitous labor, Prolonged Labor (PRIC, PROL)
Item 45 CHARACTERISTICS OF LABOR AND DELIVERY – Induction of labor, Augmentation of labor, Non-vertex presentation (INDL, AUGL, NVPR)
Item 46 METHOD OF DELIVERY- Fetal presentation at birth, Final route and method of delivery, If cesarean, was a trial of labor attempted? (PRES, ROUT, TLAB)
Item 49 BIRTHWEIGHT (BWG)
Item 50 OBSTETRIC ESTIMATION OF GESTATION (OWGEST)
Item 52 PLURALITY (PLUR)

S.10. **Denominator Exclusions** *(Brief narrative description of exclusions from the target population)*
The denominator excludes women with any of the following:

1. Gestational age at birth of less than 37 weeks or greater than 42 weeks.
2. History of a prior cesarean birth.
3. Multiple gestation.
4. Not in vertex presentation.
5. Did not attempt to have a vaginal birth by attempting labor.

S.11. **Denominator Exclusion Details** *(All information required to identify and calculate exclusions from the denominator such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
1. Gestational age at birth of less than 37 weeks or greater than 42 weeks: Exclude women whose birth certificate item number 50 (OBSTETRIC ESTIMATION OF GESTATION), processing variable: OWGEST<37. Exclude women whose birth certificate item number 50 (OBSTETRIC ESTIMATION OF GESTATION), processing variable: OWGEST>42.
2. History of a prior cesarean birth: Exclude women whose birth certificate item number 41 (RISK FACTORS IN THIS PREGNANCY), processing variable: PCES=Y.
3. Multiple gestation: Exclude women whose birth certificate item number 52 (PLURALITY), processing variable: PLUR>1.
4. Not in vertex presentation: Exclude women whose birth certificate item number 45 (CHARACTERISTICS OF LABOR AND DELIVERY), processing variable: NVPR=Y. Exclude women whose birth certificate item number 46 (METHOD OF DELIVERY), processing variable: PRES>1.
5. Did not attempt to have a vaginal birth by attempting labor: Exclude women whose birth certificate item number 46 (METHOD OF DELIVERY), processing variable: ROUT=4 AND TLAB=N UNLESS birth certificate item number 44 (ONSET OF LABOR), processing variable: PRIC=Y OR birth certificate item number 44 (ONSET OF LABOR), processing variable: PROL=Y OR birth certificate item number 45 (CHARACTERISTICS OF LABOR AND DELIVERY), processing variable: INDL=Y OR birth certificate item number 45

(CHARACTERISTICS OF LABOR AND DELIVERY), processing variable: AUGL=Y.

S.12. Stratification Details/Variables *(All information required to stratify the measure results including the stratification variables, definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format with at S.2b)*
N/A

S.13. Risk Adjustment Type (Select type. Provide specifications for risk stratification in S.12 and for statistical model in S.14-15)
Other
If other: Cohort comparison

S.14. Identify the statistical risk model method and variables *(Name the statistical method - e.g., logistic regression and list all the risk factor variables. Note - risk model development and testing should be addressed with measure testing under Scientific Acceptability)*
The statistical risk model uses a cohort comparison method derived from the concept behind logistic regression methodology. Logistic regression methodology creates an equation based on prior outcomes which is then used to predict the number of expected cesarean births in a target population. The risk adjustment is created by dividing the actual cesarean birth rate by the expected cesarean birth rate. The cohort comparison method uses the prior outcomes of similar patients to predict the number of expected cesarean births in a target population. Similar is determined by eight previously proven risk factors for cesarean birth. Those risk factors are: parity (nulliparous or multiparous), fetal size, maternal prepregnancy body mass index, maternal age, maternal height, gestational age, maternal pregnancy weight gain and type of onset of labor (spontaneous or induced).

S.15. Detailed risk model specifications *(must be in attached data dictionary/code list Excel or csv file. Also indicate if available at measure-specific URL identified in S.1.)*
Note: Risk model details (including coefficients, equations, codes with descriptors, definitions), should be provided on a separate worksheet in the suggested format in the Excel or csv file with data dictionary/code lists at S.2b.
Provided in response box S.15a

S.15a. Detailed risk model specifications (*if not provided in excel or csv file at S.2b*)
The risk adjustment is created by comparing the actual cesarean birth rate to the expected cesarean birth rate for the target population. The expected cesarean birth rate is created by first predicting the probability of cesarean birth for each woman in the target population. Each woman´s probability of cesarean birth represents her inherent risk that labor will result in a cesarean birth. Once the inherent risk for each woman in the target population is determined then the average of these inherent risks is calculated and becomes the expected cesarean birth rate for the target population.

The inherent risk for each woman is determined by finding a cohort of 100 similar women in our database of birth records. Once a cohort of at least 100 birth records of women with similar physical characteristics are found, the percentage of cesarean births found in the most recent 100 birth records is used to assign the inherent risk for that woman's birth. Similar is determined by finding birth records with the same parity (nulliparous or multiparous), onset of labor (spontaneous or induced) newborn weight plus or minus 200 grams, prepregnancy maternal body mass index plus or minus three kg/m2, maternal age plus or minus three years, maternal height plus or minus one inch, gestational age plus or minus one week, maternal weight gain plus or minus nine pounds. If the initial search does not result in 100 matching birth records within the same state then a second search is conducted ignoring the state of birth. If the second search does not result in 100 matching birth records, then the range of the physical characteristics of the birth are expanded until at least 100 matching records are found.

The Birthrisk Cesarean Birth Measure for the target population is created by multiplying the risk adjustment (actual cesarean birth rate / expected cesarean birth rate) by a constant. The average inherent risk for all births in the database that occurred in the same time frame as the target population creates the constant.

Processing variables:
1. Newborn weight: BWG from item 49
2. Prepregnancy maternal body mass index calculated from: HFT and HIN from item 31, PWGT from item 32
3. Maternal age calculated from: MDOB_YR, MDOB_MO, MDOB_DY from 8b, IDOB_YR, IDOB_MO, IDOB_DY from item 4
4. Maternal height: HFT and HIN from item 31
5. Gestational age: OWGEST from item 50
6. Maternal weight gain calculated from: DWGT from item 33, PWGT from item 32
7. Parity (nulliparous or multiparous) determined from: PLBL from item 35a, PLBD from item 35b

8. Onset of labor (spontaneous or induced): INDL from item 45

S.16. **Type of score:**
Rate/proportion
If other:

S.17. **Interpretation of Score** *(Classifies interpretation of score according to whether better quality is associated with a higher score, a lower score, a score falling within a defined interval, or a passing score)*
Better quality = Lower score

S.18. **Calculation Algorithm/Measure Logic** *(Describe the calculation of the measure score as an ordered sequence of steps including identifying the target population; exclusions; cases meeting the target process, condition, event, or outcome; aggregating data; risk adjustment; etc.)*

1. The target population is created from women who gave birth during the specified time period as determined by an existing Birth Certificate. Required data collection from each birth is obtained according to the U.S. Standard Certificate of Birth Item Number, Description and (Processing Variable(s)) as previously listed in the denominator details.

2. Women are excluded from the denominator if they gave birth prior to 37 weeks or after 42 weeks, had a history of a prior cesarean birth, had a multiple gestation, did not have a vertex presentation or did not attempt to have a vaginal birth by attempting labor as previously illustrated in the denominator exclusion details.

3. Each birth record is assigned a predicted risk of cesarean birth (inherent risk) by finding a cohort of 100 similar births in our existing database and using the number of cesarean births in the cohort as the assigned risk. Similar has been previously defined in the detailed risk model specifications.

4. The actual cesarean birth rate is determined by dividing the number of cesarean births by the number of births in the target population. The actual cesarean birth rate is determined for each obstetrical care provider and facility in the target population.

5. The expected cesarean birth rate is determined by calculating the average of the inherent risk assigned to each birth in the target population. The expected cesarean birth rate is determined for each obstetrical care provider and facility in the target population.

6. The risk adjustment is created by taking the actual cesarean birth rate and dividing it by the expected cesarean birth rate. The risk adjustment is determined for each obstetrical care provider and facility in the target population.

7. The Birthrisk Cesarean Birth Measure is created by multiplying the risk adjustment by a constant. That constant is the average inherent risk for all births occurring in the same time frame in the database as the target population and in the same state regardless of the provider or facility. The Birthrisk Cesarean Birth Measure is created for each obstetrical care provider and facility in the target population.

S.19. **Calculation Algorithm/Measure Logic Diagram URL or Attachment** *(You also may provide a diagram of the Calculation Algorithm/Measure Logic described above at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
Available at measure-specific web page URL identified in S.1

S.20. **Sampling** *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*
IF a PRO-PM, identify whether (and how) proxy responses are allowed.
N/A

S.21. **Survey/Patient-reported data** *(If measure is based on a survey, provide instructions for conducting the survey and guidance on minimum response rate.)*
IF a PRO-PM, specify calculation of response rates to be reported with performance measure results.
N/A

S.22. **Missing data** (specify how missing data are handled, e.g., imputation, delete case.)
Required for Composites and PRO-PMs.
N/A

S.23. **Data Source** *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*
If other, please describe in S.24.
Other

S.24. **Data Source or Collection Instrument** *(Identify the specific data source/data collection instrument e.g. name of database, clinical registry, collection instrument, etc.)*
IF a PRO-PM, identify the specific PROM(s); and standard methods, modes, and languages of administration.
Birth Certificate Records.

S.25. **Data Source or Collection Instrument** *(available at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
No data collection instrument provided

S.26. **Level of Analysis** *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*
Clinician : Individual, Facility

S.27. **Care Setting** *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*
Hospital/Acute Care Facility
If other:

S.28. **COMPOSITE Performance Measure** - Additional Specifications *(Use this section as needed for aggregation and weighting rules, or calculation of individual performance measures if not individually endorsed.)*

2a. Reliability – See attached Measure Testing Submission Form
2b. Validity – See attached Measure Testing Submission Form
Birthrisk_Cesarean_Birth_Measure-measure_testing_form.docx

Measure Number (*if previously endorsed*)**:** Click here to enter NQF number

Measure Title: Birthrisk Cesarean Birth Measure

Date of Submission: 2/12/2016

Type of Measure:

☐ Composite – ***STOP – use composite testing form***	☒ Outcome (*including PRO-PM*)
☐ Cost/resource	☐ Process
☐ Efficiency	☐ Structure

Instructions

- Measures must be tested for all the data sources and levels of analyses that are specified. ***If there is more than one set of data specifications or more than one level of analysis, contact NQF staff*** about how to present all the testing information in one form.
- **For all measures, sections 1, 2a2, 2b2, 2b3, and 2b5 must be completed.**
- **For outcome and resource use measures**, section **2b4** also must be completed.
- If specified for **multiple data sources/sets of specificaitons** (e.g., claims and EHRs), section **2b6** also must be completed.
- Respond to all questions as instructed with answers immediately following the question. All information on testing to demonstrate meeting the subcriteria for reliability (2a2) and validity (2b2-2b6) must be in this form. An appendix for *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 20 pages (*incuding questions/instructions;* minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.
- For information on the most updated guidance on how to address sociodemographic variables and testing in this form refer to the release notes for version 6.6 of the Measure Testing Attachment.

Note: The information provided in this form is intended to aid the Steering Committee and other stakeholders in understanding to what degree the testing results for this measure meet NQF's evaluation criteria for testing.

2a2. Reliability testing [10] demonstrates the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise. For **PRO-PMs and composite performance measures**, reliability should be demonstrated for the computed performance score.

2b2. Validity testing [11] demonstrates that the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality. For **PRO-PMs and composite performance measures**, validity should be demonstrated for the computed performance score.

2b3. Exclusions are supported by the clinical evidence; otherwise, they are supported by evidence of sufficient frequency of occurrence so that results are distorted without the exclusion; [12]
AND
If patient preference (e.g., informed decisionmaking) is a basis for exclusion, there must be evidence that the exclusion impacts performance on the measure; in such cases, the measure must be specified so that the information about patient preference and the effect on the measure is transparent (e.g., numerator category computed separately, denominator exclusion category computed separately). [13]

2b4. For outcome measures and other measures when indicated (e.g., resource use):

- **an evidence-based risk-adjustment strategy** (e.g., risk models, risk stratification) is specified; is based on patient factors (including clinical and sociodemographic factors) that influence the measured outcome and are present at start of care; [14,15] and has demonstrated adequate discrimination and calibration

OR

- rationale/data support no risk adjustment/ stratification.

2b5. Data analysis of computed measure scores demonstrates that methods for scoring and analysis of the specified measure allow for **identification of statistically significant and practically/clinically meaningful [16] differences in performance**;

OR

there is evidence of overall less-than-optimal performance.

2b6. If multiple data sources/methods are specified, there is demonstration they produce comparable results.

2b7. For **eMeasures, composites, and PRO-PMs** (or other measures susceptible to missing data), analyses identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias.

Notes

10. Reliability testing applies to both the data elements and computed measure score. Examples of reliability testing for data elements include, but are not limited to: inter-rater/abstractor or intra-rater/abstractor studies; internal consistency for multi-item scales; test-retest for survey items. Reliability testing of the measure score addresses precision of measurement (e.g., signal-to-noise).

11. Validity testing applies to both the data elements and computed measure score. Validity testing of data elements typically analyzes agreement with another authoritative source of the same information. Examples of validity testing of the measure score include, but are not limited to: testing hypotheses that the measures scores indicate quality of care, e.g., measure scores are different for groups known to have differences in quality assessed by another valid quality measure or method; correlation of measure scores with another valid indicator of quality for the specific topic; or relationship to conceptually related measures (e.g., scores on process measures to scores on outcome measures). Face validity of the measure score as a quality indicator may be adequate if accomplished through a systematic and transparent process, by identified experts, and explicitly addresses whether performance scores resulting from the measure as specified can be used to distinguish good from poor quality.

12. Examples of evidence that an exclusion distorts measure results include, but are not limited to: frequency of occurrence, variability of exclusions across providers, and sensitivity analyses with and without the exclusion.

13. Patient preference is not a clinical exception to eligibility and can be influenced by provider interventions.

14. Risk factors that influence outcomes should not be specified as exclusions

15. With large enough sample sizes, small differences that are statistically significant may or may not be practically or clinically meaningful. The substantive question may be, for example, whether a statistically significant difference of one percentage point in the percentage of patients who received smoking cessation counseling (e.g., 74 percent v. 75 percent) is clinically meaningful; or whether a statistically significant difference of $25 in cost for an episode of care (e.g., $5,000 v. $5,025) is practically meaningful. Measures with overall less-than-optimal performance may not demonstrate much variability across providers.

1. DATA/SAMPLE USED FOR ALL TESTING OF THIS MEASURE

Data used to create and validate this measure were obtained by application from the New York State Department of Health, Bureau of Biometrics and Health Statistics which contained non-identifiable information from 505,696 births certificate records from births that occurred in New York State exclusive of New York City between 2004 and 2007. Records were excluded if the birth occurred before 36 weeks 4 days or after 42 weeks 3 days of gestation (92,089), the patient had a history of a prior cesarean birth (52,833), was a multiple gestation (20,325), presentation was non-vertex (18,568) or if the patient did not attempt labor (17,797). There were 304,084 birth records that met the criteria for

inclusion of which 19,928 (6.6%) were missing or had an unusual required data element resulting in 284,156 birth records from 2,915 different obstetrical care providers working out of 109 hospitals in the initial sample.

Often the same data are used for all aspects of measure testing. In an effort to eliminate duplication, the first five questions apply to all measure testing. If there are differences by aspect of testing,(e.g., reliability vs. validity) be sure to indicate the specific differences in question 1.7.

1.1. What type of data was used for testing? (*Check all the sources of data identified in the measure specifications and data used for testing the measure. Testing must be provided for all the sources of data specified and intended for measure implementation.* ***If different data sources are used for the numerator and denominator, indicate N [numerator] or D [denominator] after the checkbox.***)

Measure Specified to Use Data From: (*must be consistent with data sources entered in S.23*)	**Measure Tested with Data From:**
☐ abstracted from paper record	☐ abstracted from paper record
☐ administrative claims	☐ administrative claims
☐ clinical database/registry	☐ clinical database/registry
☐ abstracted from electronic health record	☐ abstracted from electronic health record
☐ eMeasure (HQMF) implemented in EHRs	☐ eMeasure (HQMF) implemented in EHRs
☒ other: Birth Certificate Records	☒ other: Birth Certificate Records

1.2. If an existing dataset was used, identify the specific dataset (*the dataset used for testing must be consistent with the measure specifications for target population and healthcare entities being measured; e.g., Medicare Part A claims, Medicaid claims, other commercial insurance, nursing home MDS, home health OASIS, clinical registry*).

1.3. What are the dates of the data used in testing? 2004 - 2007

1.4. What levels of analysis were tested? (*testing must be provided for all the levels specified and intended for measure implementation, e.g., individual clinician, hospital, health plan*)

Measure Specified to Measure Performance of: (*must be consistent with levels entered in item S.26*)	**Measure Tested at Level of:**
☒ individual clinician	☒ individual clinician
☐ group/practice	☐ group/practice
☒ hospital/facility/agency	☒ hospital/facility/agency
☐ health plan	☐ health plan
☐ other: Click here to describe	☐ other: Click here to describe

1.5. How many and which measured entities were included in the testing and analysis (by level of analysis and data source)? (*identify the number and descriptive characteristics of measured entities included in the analysis (e.g., size, location, type); if a sample was used, describe how entities were selected for inclusion in the sample*)

There were 2,915 obstetrical care providers working out of 109 facilities in the data obtained from New York State. This data included all obstetrical care providers and facilities in New York State exclusive of those within New York City.

1.6. How many and which patients were included in the testing and analysis (by level of analysis and data source)? (*identify the number and descriptive characteristics of patients included in the analysis (e.g., age, sex, race, diagnosis); if a sample was used, describe how patients were selected for inclusion in the sample*)

Of the 505,696 birth records from 2004-2007 that were obtained, all of the patients who attempted labor without a history of a prior cesarean birth and gave birth to a single baby in vertex presentation between 36 weeks 4 days and 42 weeks 3 days of gestation were included in the testing and analysis There were 304,084 birth records that met the criteria for inclusion of which 19,928 (6.6%) were missing or had an unusual required data element resulting in 284,156

birth records from 2,915 different obstetrical care providers working out of 109 hospitals in the initial sample. The number and characteristics of the data file are provided in the following three tables.

Year	2004	2005	2006	2007	Total
Total records	**127,489**	**125,297**	**126,446**	**126,464**	**505,696**
Non term gestations	23,296	21,871	23,609	23,313	92,089
Prior cesarean delivery	12,627	12,604	12,625	14,977	52,833
Multiple gestations	5,204	4,813	5,028	5,280	20,325
Non vertex presentations	5,366	4,648	4,613	3,941	18,568
Did not labor	3,136	4,753	5,641	4,267	17,797
Unknown delivery type	394	97	110	31	632
Bad facility code	551	569	583	692	2,395
Missing or unusual information	4,138	3,935	4,882	3,946	16,901
Eligible for analysis	**72,777**	**72,007**	**69,355**	**70,017**	**284,156**
Nulliparous records	33,158	32,967	32,118	33,113	131,356
Spontaneous nulliparous records	24,508	23,464	22,795	23,146	93,913
Induced nulliparous records	8,650	9,503	9,323	9,967	37,443
Multiparous records	39,619	39,040	37,237	36,904	152,800
Spontaneous multiparous records	30,995	29,446	27,794	27,619	115,854
Induced multiparous records	8,624	9,594	9,443	9,285	36,946
Number of Hospitals	**109**	**104**	**101**	**101**	**109**
Number of Providers	**2,107**	**2,031**	**1,896**	**1,881**	**2,915**

Nulliparous (N=131,356)		
	n (Spont./Induced)	**Total (%)**
Height (inches)		
<60	4,052 (2,983/1,069)	3.08
60-62	31,889 (23,180/8,709)	24.28
63-65	54,381 (39,026/15,355)	41.40
66-68	33,525 (23,555/9,970)	25.52
>68	7,509 (5,169/2340)	5.72
Fetal Weight (grams)		
<2500	3,302 (2,104/1,198)	2.51
2500-2900	14,922 (10,890/4,032)	11.36
2901-3300	38,981 (29,345/9,636)	29.68
3301-3700	43,698 (31,458/12,240)	33.27
3701-4100	22,955 (15,421/7,534)	17.48
4101-4500	6,314 (3,973/2,341)	4.81
>4500	1,184 (722/462)	0.90
Prepregnancy BMI (kg/m^2)		
<20	12,187 (9,540/2,647)	9.28
20-26	78,478 (58,367/20,111)	59.74

27-33	29,033 (19,336/9,697)	22.10
34-40	8,609 (5,051/3,558)	6.55
>40	3,049 (1,619/1,430)	2.32
Age (years)		
<20	18,951 (14,111/4,840)	14.43
20-26	47,541 (34,252/13,289)	36.19
27-33	47,735 (33,917/13,818)	36.34
34-40	15,902 (10,864/5,038)	12.11
>40	1,227 (769/458)	0.93
Gestational Age (weeks)		
37	5,207 (3,791/1,416)	3.96
38	15,937 (11,829/4,108)	12.13
39	30,198 (22,737/7,461)	22.99
40	39,505 (29,602/9,903)	30.07
41	30,206 (19,629/10,577)	23.00
42	10,303 (6,325/3,978)	7.84
Weight Gain (lbs.)		
<11	5,695 (3,965/1,730)	4.34
11-29	42,489 (31,202/11,287)	32.35
30-48	64,060 (46,223/17,837)	48.77
>48	19,112 (12,523/6,589)	14.55

Multiparous (N=152,800)		
	n (Spont./Induced)	**Total (%)**
Height (inches)		
<60	4,522 (3718/804)	2.96
60-62	35,672 (28,120/7552)	23.35
63-65	62,335 (47,227/15,108)	40.80
66-68	40,750 (30,047/10,703)	26.67
>68	9,521 (6,742/2,779)	6.23
Fetal Weight (grams)		
<2500	2,556 (1,792/764)	1.67
2500-2900	13,041 (9,978/3,063)	8.53
2901-3300	39,845 (31,180/8,665)	26.08
3301-3700	52,356 (40,020/12,336)	34.26
3701-4100	32,225 (23,864/8,361)	21.09
4101-4500	10,509 (7,398/3,111)	6.88
>4500	2,268 (1,622/646)	1.48
Prepregnancy BMI (kg/m^2)		
<20	11,752 (9,532/2,220)	7.69
20-26	85,516 (66,818/18,698)	55.97
27-33	39,661 (29,097/10,564)	25.96
34-40	11,919 (8,006/3,913)	7.80
>40	3,952 (2,401/1,551)	2.59

Age (years)		
<20	2,748 (2,244/504)	1.80
20-26	40,564 (31,556/9,008)	26.55
27-33	64,467 (48,618/15,849)	42.19
34-40	41,198 (30,719/10,479)	26.96
>40	3,823 (2,717/1,106)	2.50
Gestational Age (weeks)		
37	6,689 (5,254/1,435)	4.38
38	22,503 (17,285/5,218)	14.73
39	42,574 (32,413/10,161)	27.86
40	45,455 (35,466/9,989)	29.75
41	26,998 (19,347/7,651)	17.67
42	8,581 (6,089/2,492)	5.62
Weight Gain (lbs.)		
<11	11,530 (8,587/2,943)	7.55
11-29	64,596 (49,602/14,994)	42.27
30-48	64,709 (49,230/15,479)	42.35
>48	11,965 (8,435/3,530)	7.83

1.7. If there are differences in the data or sample used for different aspects of testing (e.g., reliability, validity, exclusions, risk adjustment), identify how the data or sample are different for each aspect of testing reported below.

No differences

1.8 What were the patient-level sociodemographic (SDS) variables that were available and analyzed in the data or sample used? For example, patient-reported data (e.g., income, education, language), proxy variables when SDS data are not collected from each patient (e.g. census tract), or patient community characteristics (e.g. percent vacant housing, crime rate).

Sociodemographic variables were not collected or analyzed.

2a2. RELIABILITY TESTING

Note*: If accuracy/correctness (validity) of data elements was empirically tested, separate reliability testing of data elements is not required – in 2a2.1 check critical data elements; in 2a2.2 enter "see section 2b2 for validity testing of data elements"; and skip 2a2.3 and 2a2.4.*

2a2.1. What level of reliability testing was conducted? (*may be one or both levels*)

☒ **Critical data elements used in the measure** (*e.g., inter-abstractor reliability; data element reliability must address ALL critical data elements*)

☒ **Performance measure score** (e.g., *signal-to-noise analysis*)

2a2.2. For each level checked above, describe the method of reliability testing and what it tests (*describe the steps—do not just name a method; what type of error does it test; what statistical analysis was used*)

In 2006 Northam and Knapp published a review of twenty four primary research studies of U.S. birth certificates that involved validity or reliability assessment. They reported that: "Reliability is admittedly difficult to evaluate for birth certificate data. The dichotomous data are not amenable to traditional internal consistency evaluation because birth certificates provide data about individual variables rather than dimensions of a single construct, and they use a single item for each variable. Because birth certificate data reflect a cross-sectional assessment of each birth and are

concerned with variables such as birthweight, obstetric care, and the presence of infant abnormalities that are reported at one time, they are also not amenable to an investigation of temporal stability. Thus, both internal consistency and stability assessments are not feasible methods of evaluating the reliability of birth certificate data."

See section 2b2 for validity testing of data elements

A signal to noise analysis for the hospital performance measure score was conducted to provide reliability testing for the performance measure score. MS group is the signal (2.532). MS error is the noise (27.910). This will test if the between-group variance (signal) is comparable to the within-group variance (noise). Signal-to-noise ratio was calculated as follows: 2.532/27.910 = 0.09072

A signal to noise analysis for the provider performance measure score was conducted to provide reliability testing for the performance measure score. MS group is the signal (777.361). MS error is the noise (709.237). This will test if the between-group variance (signal) is comparable to the within-group variance (noise). Signal-to-noise ratio was calculated as follows: 777.361/709.237= 1.096053

2a2.3. For each level of testing checked above, what were the statistical results from reliability testing? (e.*g., percent agreement and kappa for the critical data elements; distribution of reliability statistics from a signal-to-noise analysis*)

Hospital Performance Measure Score:

ANOVA

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	5.063	2	2.532	.091	.913
Within Groups	8456.786	303	27.910		
Total	8461.849	305			

Provider Performance Measure Score:

ANOVA

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	1554.722	2	777.361	1.096	.334
Within Groups	4118540.381	5807	709.237		
Total	4120095.102	5809			

2a2.4 What is your interpretation of the results in terms of demonstrating reliability? (i.*e., what do the results mean and what are the norms for the test conducted?*)

A ratio higher than 1:1 (greater than 0 dB) indicates more signal than noise. Thus, the between-group variance (signal) is comparable to the within-group variance. The Cronbach's alpha was .87 that was greater than .70 (Nyunnally, 1978) indicating high reliability.

2b2. VALIDITY TESTING

2b2.1. What level of validity testing was conducted? (*may be one or both levels*)

☒ **Critical data elements** (*data element validity must address ALL critical data elements*)

☒ **Performance measure score**

☒ **Empirical validity testing**

☐ **Systematic assessment of face validity of <u>performance measure score</u> as an indicator** of quality or resource use (*i.e., is an accurate reflection of performance on quality or resource use and can distinguish good from poor performance*)

2b2.2. For each level of testing checked above, describe the method of validity testing and what it tests (*describe the steps—do not just name a method; what was tested, e.g., accuracy of data elements compared to authoritative source, relationship to another measure as expected; what statistical analysis was used)*

Northam and Knapp reported that "The 1st U.S. Standard Certificate of Live Birth was developed in 1900 (Gould, 1999), and birth certificates have included data on maternal and infant variables for all live births in the United States and its territories since 1939 (Lilienfeld, Parkhurst, Patton, & Schlesinger, 1951)." Despite the long history regarding the use of the birth certificate, validity testing of the data elements provided by birth certificate data remains elusive with most studies regarding the validity of birth certificate data elements being conducted prior the 2003 revision of the U.S. birth certificate. Recently Martin et al. published a National Vital Statistics Report in 2013 comparing the data elements in 600 birth certificate records in one state and 495 birth certificate records in another to the information recorded in the medical record. Each data element in the medical record was compared to the birth certificate data element and the percentage of exact matches was recorded.

Performance measure score validity testing was conducted using construct validity for the hospital performance measure score and criterion validity for the provider performance measure score. The type of criterion validity that was used was predictive validity.

2b2.3. What were the statistical results from validity testing? (*e.g., correlation; t-test*)

Validity of the data elements was examined in the National Vital Statistics Report from 2013 which compared many but not all of the birth certificate data elements to the results found in the medical record and reported the findings based on the percentage of exact agreement as follows:

	Percent Agreement	
U. S. Birth Certificate Item	State A	State B
Number of previous live births now living	96.0%	96.1%
Number of previous live births now dead	99.2%	97.9%
Number of previous cesarean deliveries	95.3%	92.5%
Obstetric estimate of gestation at delivery (exact weeks)	91.6%	67.4%
Obstetric estimate of gestation (within 2 weeks)	99.7%	98.1%
Birthweight (exact grams)	90.0%	90.7%
Birthweight within 500 grams	99.7%	99.4%
Induction of labor	89.7%	83.0%
Cephalic Presentation	92.0%	93.2%
Breech Presentation	98.2%	98.0%
Cesarean Birth	98.8%	97.2%
Trial of labor attempted	89.0%	84.6%

An older report from Querec (1980) comparing birth certificate data with maternal survey data revealed that excellent agreement (90% or better) was noted for maternal age and plurality.

The performance measure score validity for the hospital performance measure score revealed that the Cronbach's alpha was .87. The performance measure score predictive validity for the provider performance measure score revealed a regression line with an equation of y = 1.017x + 1.4964 and an R-squared value of 0.9802.

2b2.4. What is your interpretation of the results in terms of demonstrating validity? (i*.e., what do the results mean and what are the norms for the test conducted?*)

2a1. The data elements that have been studied were demonstrated to have high validity when compared to the medical record with the exception of Induction of Labor and Trial of labor which were deemed to have substantial validity. There was a discrepancy in Obstetric estimate of gestation at delivery (exact weeks) between the two states with state B only having moderate validity.

2a2.

2a3. Unfortunately not all of the data elements that are relevant to this measure have been tested for validity. The data elements of Mother's Height, Mother's Prepregnancy Weight and Mother's Weight at Delivery were not addressed any in of the studies. Further testing will be required to determine if validity of these data elements is similar to the validity of the tested data elements.

2a4.

2a5. The Cronbach's alpha was 0.87, which was greater than .70 (Nyunnally, 1978). High reliability implies high construct validity. The alpha coefficient of the performance measure score was greater than .70 (Nyunnally, 1978). Thus, the hospital performance measure score has high construct validity.

2a6.

2a7. The performance measure score predictive validity testing reveals an excellent correlation between the actual cesarean birth rate to the predicted cesarean birth rate using the provider performance measure score.

2b3. EXCLUSIONS ANALYSIS

NA ☒ no exclusions — *skip to section 2b4*

2b3.1. Describe the method of testing exclusions and what it tests (*describe the steps—do not just name a method; what was tested, e.g., whether exclusions affect overall performance scores; what statistical analysis was used*)

2b3.2. What were the statistical results from testing exclusions? (*include overall number and percentage of individuals excluded, frequency distribution of exclusions across measured entities, and impact on performance measure scores*)

2b3.3. What is your interpretation of the results in terms of demonstrating that exclusions are needed to prevent unfair distortion of performance results? (*i.e., the value outweighs the burden of increased data collection and analysis. Note: **If patient preference is an exclusion**, the measure must be specified so that the effect on the performance score is transparent, e.g., scores with and without exclusion*)

2b4. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES

If not an intermediate or health outcome, or PRO-PM, or resource use measure, skip to section 2b5.

2b4.1. What method of controlling for differences in case mix is used?

☐ **No risk adjustment or stratification**

☐ **Statistical risk model with** Click here to enter number of factors **risk factors**

☐ **Stratification by** Click here to enter number of categories **risk categories**

☒ **Other, Cohort comparison**

2b4.2. If an outcome or resource use measure is not risk adjusted or stratified, provide rationale and analyses to demonstrate that controlling for differences in patient characteristics (case mix) is not needed to achieve fair comparisons across measured entities.

N/A

2b4.3. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors (clinical factors or sociodemographic factors) used in the statistical risk model or for stratification by risk (*e.g., potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of p<0.10; correlation of x or higher; patient factors should be present at the start of care*)

The factors selected for the risk adjustment were: parity (nulliparous or multiparous), newborn weight, prepregnancy maternal body mass index, maternal age, maternal height, gestational age, maternal weight gain, and the type of onset of labor (spontaneous or induced). These factors were selected because they are objective, are typically measured in every pregnancy and have been previously proven to significantly affect the risk that labor will result in a cesarean birth.

The selected factors were confirmed to significantly affect the risk that labor will result in a cesarean birth in the following manner. Parity and onset of labor were analyzed by z-test. Newborn weight, prepregnancy maternal body mass index, maternal age, maternal height, gestational age and maternal weight gain were analyzed using multivariate logistic regression analysis.

2b4.4a. What were the statistical results of the analyses used to select risk factors?

Analysis of the 284,156 birth records revealed that 32,540 (24.77%) of the 131,356 nulliparous women had a cesarean birth and 7,679 (5.03%) of the 152,800 multiparous women had a cesarean birth. The cesarean birth rate in nulliparous women was significantly higher than in multiparous women (p<=0.001). Analysis of the 131,356 nulliparous birth records revealed that 13,558 (36.21%) of the 37,443 women whose labor was induced had a cesarean birth and 18,982 (20.12%) of the 93,913 women whose labor was spontaneous had a cesarean birth. The cesarean birth rate for nulliparous women whose labor was induced was significantly higher than in nulliparous women whose labor was spontaneous (p<=0.001). Analysis of the 152,800 multiparous birth records revealed that 2,749 (7.44%) of the 36,946 women whose labor was induced had a cesarean birth and 4,930 (4.26%) of the 115,854 women whose labor was spontaneous had a cesarean birth. The cesarean birth rate for multiparous women whose labor was induced was significantly higher than in multiparous women whose labor was spontaneous (p<=0.001).

Both newborn weight and maternal weight gain did not exhibit a linear progression throughout the reported ranges. Increasing newborn weight when over 2900 grams, increasing prepregnancy body mass index, increasing maternal age, decreasing maternal height, decreasing newborn weight when under 2900 grams, longer gestations, increasing maternal weight gain when over 25 pounds and decreasing maternal weight gain when under 25 pounds significantly affected the risk for cesarean birth for both nulliparous and multiparous women (p<=0.001). The effect of each physical characteristic on the rate of cesarean birth for the 284,156 women is illustrated in figures 1 through 6 and in the two tables that follow.

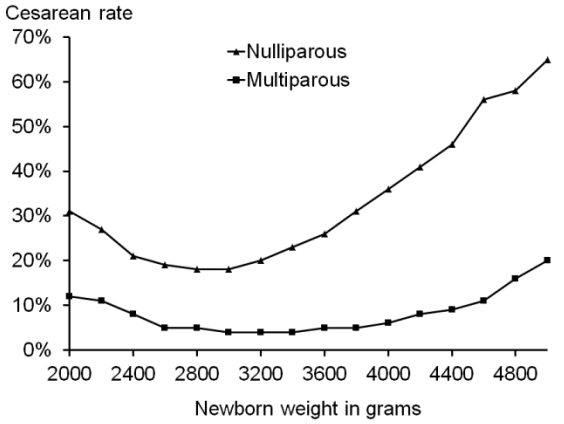


Figure 1 – Effect of newborn weight

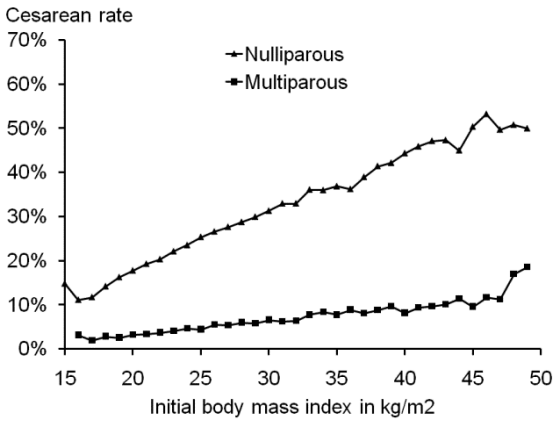


Figure 2 – Effect of prepregnancy body mass index

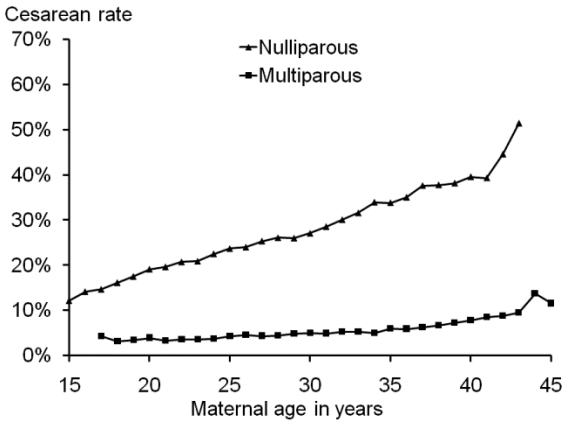


Figure 3 – Effect of maternal age

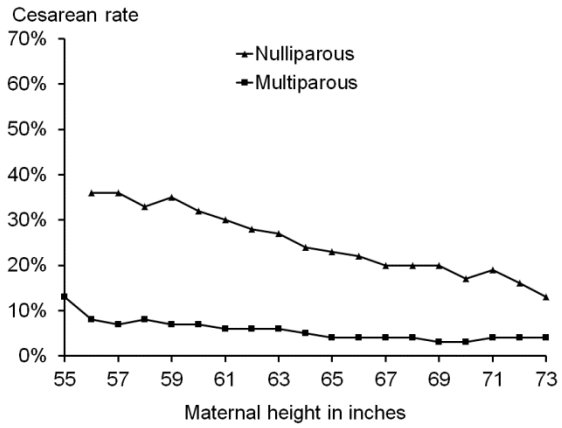


Figure 4 – Effect of maternal height

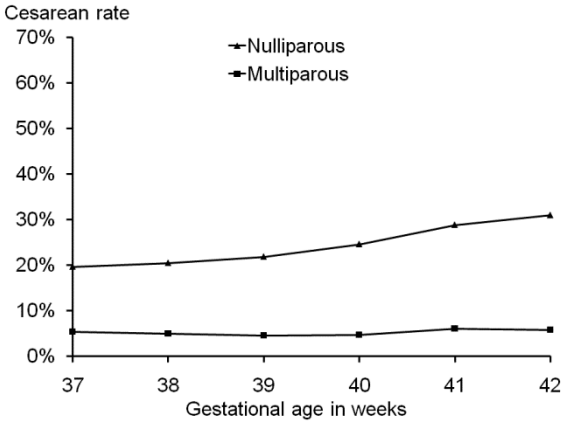


Figure 5 – Effect of gestational age

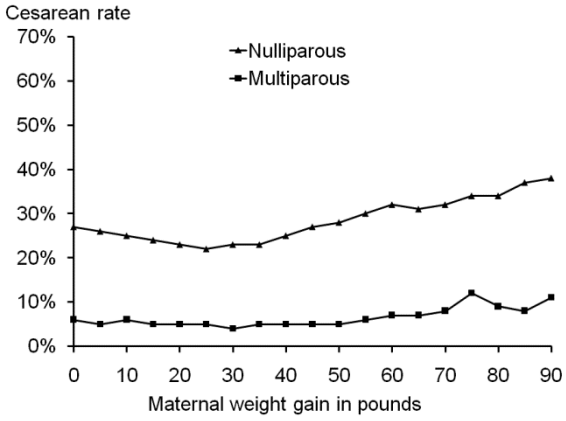


Figure 6 – Effect of maternal weight gain

Adjusted odds ratio progression of cesarean birth for nulliparous women:

Physical characteristic	Step Size	Odds Ratio	95% Confidence Interval
Newborn weight when > 2900 g	Every 200 g larger	1.331	1.310-1.352

Prepregnancy maternal body mass index	Every 3 kg/m^2 increase	1.210	1.204-1.219
Maternal age	Every 3 years older	1.195	1.186-1.201
Maternal height	Every inch shorter	1.124	1.119-1.128
Newborn weight when <= 2900 g	Every 200 g smaller	1.122	1.107-1.134
Gestational age	Every week longer	1.100	1.088-1.113
Maternal weight gain when > 25 lbs	Every 5 lbs more	1.078	1.070-1.085
Maternal weight gain when <= 25 lbs	Every 5 lbs less	1.020	1.005-1.035

Adjusted odds ratio progression of cesarean birth for multiparous women:

Physical characteristic	Step Size	Odds Ratio	95% Confidence Interval
Prepregnancy body mass index	Every 3 kg/m^2 increase	1.201	1.189-1.210
Newborn weight when <= 2900 g	Every 200 g smaller	1.155	1.140-1.165
Maternal age	Every 3 years older	1.153	1.141-1.165
Newborn weight when > 2900 g	Every 200 g larger	1.149	1.127-1.171
Maternal height	Every inch shorter	1.091	1.083-1.099
Maternal weight gain when > 25 lbs	Every 5 lbs more	1.085	1.075-1.095
Gestational age	Every week longer	1.035	1.014-1.055
Maternal weight gain when <= 25 lbs	Every 5 lbs less	1.030	1.005-1.055

2b4.4b. Describe the analyses and interpretation resulting in the decision to select SDS factors (e.g. prevalence of the factor across measured entities, empirical association with the outcome, contribution of unique variation in the outcome, assessment of between-unit effects and within-unit effects)

SDS factors were not selected.

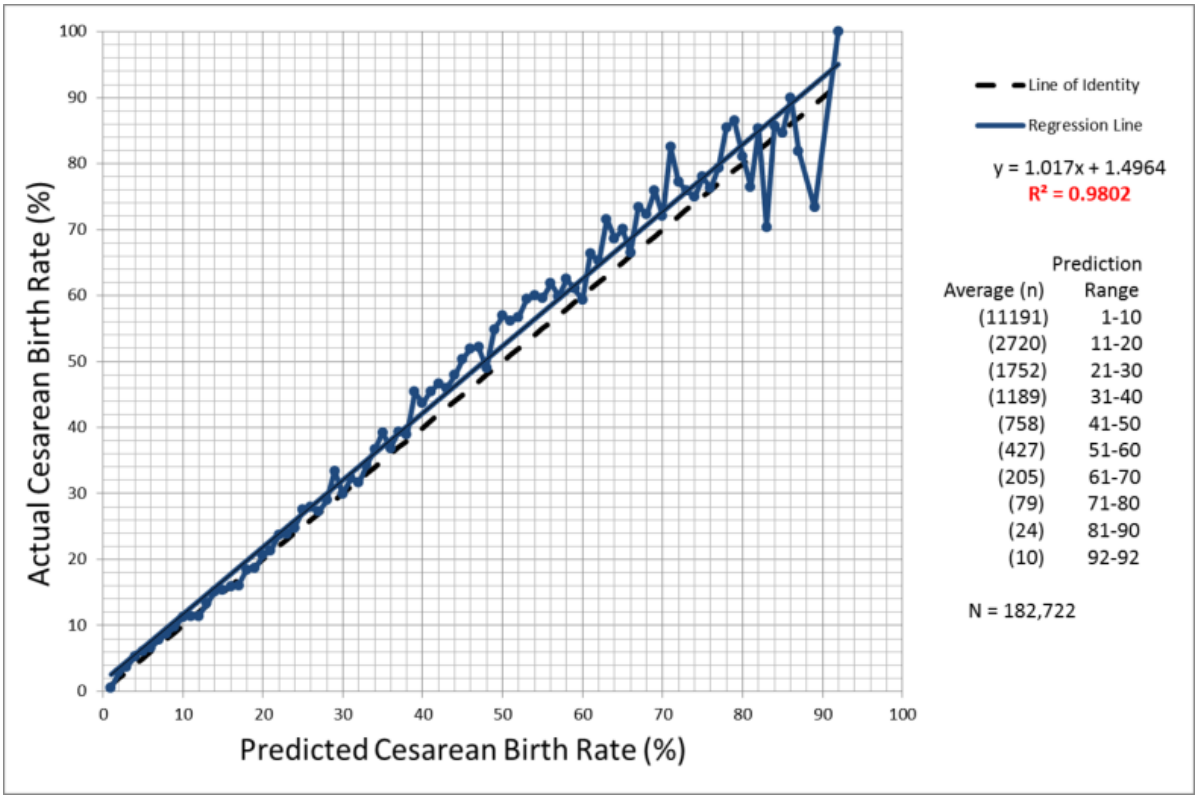
2b4.5. Describe the method of testing/analysis used to develop and validate the adequacy of the statistical model or stratification approach (*describe the steps—do not just name a method; what statistical analysis was used*)

Validation of the accuracy of the risk adjustment is achieved by using the risk adjustment to predict the outcomes for women whose births were not used to create the risk adjustment. The inherent risk previously assigned to each of 211,379 test records (those from 2005 to 2007) along with the obstetrical care provider's risk adjustment at the time of that birth is used to calculate the predicted outcome for that test record. Predictions were limited to a range from 1% to 99%. However, if the obstetrical care provider did not yet have ten birth records in the database then that test record was not assigned a predicted risk.

This resulted in 182,757 birth records for which a predicted risk of cesarean birth was calculated. The 182,757 records were grouped according to their predicted risk from 1% to 99%. Ten of these groups had less than ten test records in the group and were not used in the comparison. The predicted risk groups that were omitted from the comparison were 88%, 90%, 91% and 93% to 99% inclusive. This resulted in 182,722 records for comparison. The actual cesarean birth rate for each group was compared to the predicted cesarean birth rate for each group.

Provide the statistical results from testing the approach to controlling for differences in patient characteristics (case mix) below.
If stratified, skip to 2b4.9

2b4.6. Statistical Risk Model Discrimination Statistics (*e.g., c-statistic, R-squared*)**:**

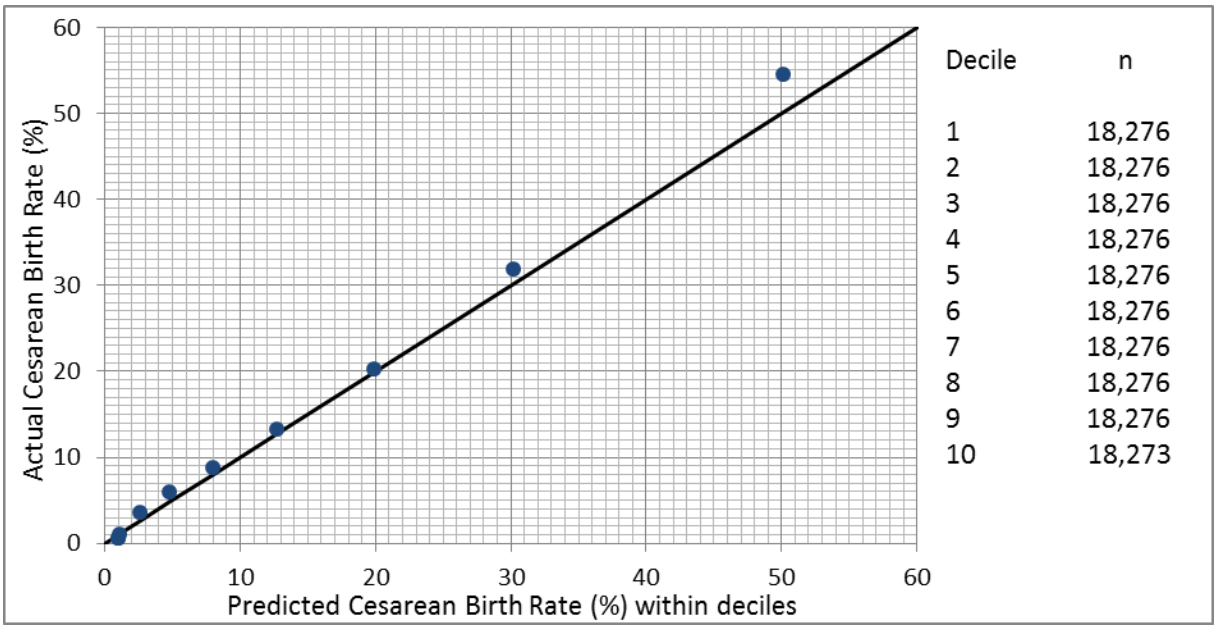


The comparison of the actual cesarean birth rate to the predicted cesarean birth rate resulted in a regression line with an equation of y = 1.017x + 1.4964 and an R-squared of 0.9802.

2b4.7. Statistical Risk Model Calibration Statistics (*e.g., Hosmer-Lemeshow statistic*):

The Hosmer-Lemeshow test statistic is 348.34 (df=8; p-value=0.00).

2b4.8. Statistical Risk Model Calibration – Risk decile plots or calibration curves:



2b4.9. Results of Risk Stratification Analysis:

2b4.10. What is your interpretation of the results in terms of demonstrating adequacy of controlling for differences in patient characteristics (case mix)? (i.*e., what do the results mean and what are the norms for the test conducted*)

Validation of the risk adjustment was obtained by predicting the outcome for births that were not included in assigning each obstetrical care provider's risk adjustment. The large number of predictions in the validation allowed for a percentage point by percentage point analysis which revealed an R-squared value of 0.9802. This high R-squared value reveals the accuracy of both the inherent risk assigned to each birth and the risk adjustment assigned to the obstetrical care provider.

In the validation, the obstetrical care provider's risk adjustment from their previous 100 births is used to assign the predicted risk to each subsequent birth. Therefore, if the cesarean birth rate is changing over time the predicted outcomes may underestimate or overestimate the actual outcomes. The cesarean birth rate increased from 13.45% in 2004 to 14.12%, 14.19% and 14.89% in the following three years. As expected with an increasing cesarean birth rate the predictions are slightly underestimated. The large sample size and the underestimation due to the increasing cesarean birth rate may explain the failure of the Hosmer-Lemeshow test. A changing cesarean birth rate only affects the validation of the risk adjustment and not the ability to compare cesarean birth measures between entities for a specific period of time.

2b4.11. Optional Additional Testing for Risk Adjustment (*not required, but would provide additional support of adequacy of risk model, e.g., testing of risk model in another data set; sensitivity analysis for missing data; other methods that were assessed*)

2b5. IDENTIFICATION OF STATISTICALLY SIGNIFICANT & MEANINGFUL DIFFERENCES IN PERFORMANCE

2b5.1. Describe the method for determining if statistically significant and clinically/practically meaningful differences in performance measure scores among the measured entities can be identified (*describe the steps—do not just name a method; what statistical analysis was used? Do not just repeat the information provided related to performance gap in 1b)*

The Birthrisk Cesarean Birth Measure provides the ability to assess differences in performance at several levels. The obstetrical care provider can compare themselves to the other providers in their facility, county, state or across the nation. A facility can compare themselves to the other facilities in their county, state or across the nation. Healthcare administrators can compare results at the level of the obstetrical care provider, facility, county, state or across the nation. A simple z-test is used to create a 95% confidence interval to determine if the selected cesarean birth measure is statistically significantly different than that of another entity or that of the average in the facility, county, state or nation.

2b5.2. What were the statistical results from testing the ability to identify statistically significant and/or clinically/practically meaningful differences in performance measure scores across measured entities? (e.g., *number and percentage of entities with scores that were statistically significantly different from mean or some benchmark, different from expected; how was meaningful difference defined*)

There were 211,379 birth records (2005-2007) from 2,542 different obstetrical care providers working out of 105 different facilities in New York State that were used to assess performance in the initial sample.

Comparing the 2,542 obstetrical care providers to the average cesarean birth measure for the state, there were 1,802 (70.9%) obstetrical care providers that had a cesarean birth measure that was not significantly different than the average and they represented 55% of the total births. There were 398 (15.6%) obstetrical care providers that had a cesarean birth measure that was significantly better than the average and they represented 26% of the total births. There were 342 (13.5%) obstetrical care providers that had a cesarean birth measure that was significantly worse than the average and they represented 19% of the total births.

Comparing the 105 facilities to the average cesarean birth measure for the state, there were 42 (40%) facilities that had a cesarean birth measure that was not significantly different than the average and they represented 28.6% of the total births. There were 30 (29%) facilities that had a cesarean birth measure that was significantly better than the average and they represented 38.0% of the total births. There were 33 (31%) facilities that had a cesarean birth measure that was significantly worse than the average and they represented 33.4% of the total births.

Comparing the 86 obstetrical care providers who work in the New York State facility with the most births (9,364) to the average cesarean birth measure for the state, there were 72 (84%) obstetrical care providers that had a cesarean birth measure that was not significantly different than the average and they represented 71% of the total births. There were 5 (9%) obstetrical care providers that had a cesarean birth measure that was significantly better than the average and they represented 9% of the total births. There were 9 (10%) obstetrical care providers that had a cesarean birth measure that was significantly worse than the average and they represented 18% of the total births. This facility was one of the 33 that had a cesarean birth measure that was worse than the average for the state.

2b5.3. What is your interpretation of the results in terms of demonstrating the ability to identify statistically significant and/or clinically/practically meaningful differences in performance across measured entities? (i.*e., what do the results mean in terms of statistical and meaningful differences?*)

The decision to perform a cesarean birth is made by obstetrical care providers and not by facilities. Therefore, the level of analysis required to improve outcome must be down to the level of the obstetrical care provider. For example, in the analysis of the facility with the most births it is the 9 obstetrical care providers in that facility who are performing 18% of the births that results in that facility being one of the 33 facilities that was found to have a cesarean birth measure that was significantly worse than the average for the state.

The Birthrisk Cesarean Birth Measure uses a simple z-test to create statistically significant comparisons between entities. The ability to report an obstetrical care provider's cesarean birth measure as a comparison to the average in their facility, county, state or in the nation with a *p* value gives obstetrical care providers results in a manner with which they are familiar. Obstetrical care providers and facilities would have fewer objections to public reporting of their cesarean birth measure if the reporting is limited to providing results as better than average, average or worse than average based on a 95% confidence interval.

Reporting results based on a 95% confidence interval would also allow for an acceptable plan for value based incentive/disincentive payments. Paying providers or facilities more if they are significantly better than average and paying providers or facilities less if they are significantly worse than average will provide the correct incentive for the management of laboring patients because providers who perform more cesarean births than expected actually work less than those who perform less cesarean births than expected. This is because it is less work for the obstetrical care provider to end a labor by proceeding to a cesarean birth than it is to allow the labor to continue.

2b6. COMPARABILITY OF PERFORMANCE SCORES WHEN MORE THAN ONE SET OF SPECIFICATIONS
***If only one set of specifications, this section can be skipped*.**

Note: *This item is directed to measures that are risk-adjusted (with or without SDS factors)* ***OR*** *to measures with more than one set of specifications/instructions (e.g., one set of specifications for how to identify and compute the measure from medical record abstraction and a different set of specifications for claims or eMeasures). It does not apply to measures that use more than one source of data in one set of specifications/instructions (e.g., claims data to identify the denominator and medical record abstraction for the numerator).* ***Comparability is not required when comparing performance scores with and without SDS factors in the risk adjustment model. However, if comparability is not demonstrated for measures with more than one set of specifications/instructions, the different specifications (e.g., for medical records vs. claims) should be submitted as separate measures.***

2b6.1. Describe the method of testing conducted to compare performance scores for the same entities across the different data sources/specifications (*describe the steps—do not just name a method; what statistical analysis was used*)

2b6.2. What were the statistical results from testing comparability of performance scores for the same entities when using different data sources/specifications? (*e.g., correlation, rank order*)

2b6.3. What is your interpretation of the results in terms of the differences in performance measure scores for the same entities across the different data sources/specifications? (i.*e., what do the results mean and what are the norms for the test conducted*)

2b7. MISSING DATA ANALYSIS AND MINIMIZING BIAS

2b7.1. Describe the method of testing conducted to identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias (*describe the steps—do not just name a method; what statistical analysis was used*)

Birth records in the initial sample were analyzed using IBM SPSS Statistical software version 23 and Little's MCAR test was performed. Analysis included unconditional testing to find records that were missing any of the data elements or had data elements that were outside of the usual accepted range. After the unconditional tests of known requirements were performed, critical data elements were subjected to conditional tests to assure that the logical relationships exist and were appropriate. Listwise deletion of records with missing or unusual data elements was performed. The dataset of deleted records was compared to the remaining data sample. The distribution of deleted records by providers was analyzed. Bias is minimized by confirming that deleted records were similar to retained records. The ability to re-abstract data in future samples will result in less missing values which in turn will further reduce any potential bias.

2b7.2. What is the overall frequency of missing data, the distribution of missing data across providers, and the results from testing related to missing data? (*e.g., results of sensitivity analysis of the effect of various rules for missing data/nonresponse; if no empirical sensitivity analysis, identify the approaches for handling missing data that were considered and pros and cons of each*)

There were 224,142 records that met the criteria for inclusion in order to assess performance. Of these records there were 12,763 (5.69%) that were missing one or more of the critical data elements or had unusual data. Little's MCAR test revealed that the missing values were not random. Analysis of the data records that were removed from the data sample revealed that 1,229 providers had at least one record deleted. For the 1,229 providers who had records deleted the average percentage of deleted records was 9.85%. Analysis of the means for the data elements revealed:

	Parity - Nulliparous	Induced	Baby Wt.	Prepreg Wt.	Postpreg Wt.	Age	Height	Weeks
Deleted Records	43.79%	20.01%	3406	149.99	180.79	28.09	64.18	39.71
Retained Records	46.46%	27.02%	3421	150.48	182.07	28.31	64.24	39.68

The cesarean birth rate of the deleted records was 14.1% which was not statistically significantly different from the 14.4% cesarean birth rate for the 211,379 birth records used to assess performance.

Missing data was handled with Listwise deletion. Imputation methods were specifically not used as the performance measure is dependent on the unique combination of physical characteristics of the mother and the size of her baby. Improving data collection through re-abstraction will improve results.

2b7.3. What is your interpretation of the results in terms of demonstrating that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias**?** (i.*e., what do the results mean in terms of supporting the selected approach for missing data and what are the norms for the test conducted; if no empirical analysis, provide rationale for the selected approach for missing data*)

The overall amount of deleted records is small and analysis reveals no significant difference in the physical characteristics of the deleted and retained records. The small difference in parity and induction of deleted records is unlikely to introduce bias because of the cohort comparison method that is used to create the risk adjustment for the performance measure. Since reducing the number of missing values provides the best method for reducing any potential bias, future data collections will benefit from re-abstraction, mandatory data element completion and verification of unusual data ele

3. Feasibility

Extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

3a. Byproduct of Care Processes

For clinical measures, the required data elements are routinely generated and used during care delivery (e.g., blood pressure, lab test, diagnosis, medication order).

3a.1. **Data Elements Generated as Byproduct of Care Processes.**

Generated or collected by and used by healthcare personnel during the provision of care (e.g., blood pressure, lab value, diagnosis, depression score)

If other:

3b. Electronic Sources

The required data elements are available in electronic health records or other electronic sources. If the required data are not in electronic health records or existing electronic sources, a credible, near-term path to electronic collection is specified.

3b.1. **To what extent are the specified data elements available electronically in defined fields?** *(i.e., data elements that are needed to compute the performance measure score are in defined, computer-readable fields)*

ALL data elements are in defined fields in a combination of electronic sources

3b.2. **If ALL the data elements needed to compute the performance measure score are not from electronic sources, specify a credible, near-term path to electronic capture, OR provide a rationale for using other than electronic sources.**

3b.3. **If this is an eMeasure, provide a summary of the feasibility assessment in an attached file or make available at a measure-specific URL.**

Attachment:

3c. Data Collection Strategy

Demonstration that the data collection strategy (e.g., source, timing, frequency, sampling, patient confidentiality, costs associated with fees/licensing of proprietary measures) can be implemented (e.g., already in operational use, or testing demonstrates that it is ready to put into operational use). For eMeasures, a feasibility assessment addresses the data elements and measure logic and demonstrates the eMeasure can be implemented or feasibility concerns can be adequately addressed.

3c.1. **Describe what you have learned/modified as a result of testing and/or operational use of the measure regarding data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, time and cost of data collection, other feasibility/implementation issues.**

IF a PRO-PM, consider implications for both individuals providing PROM data (patients, service recipients, respondents) and those whose performance is being measured.

The goal is to have each entity provide all of their data electronically. The data can be digitally extracted from the facility from the information used to report data for the U.S. Standard Certificate of Birth. Ideally, the data will be provided by a software bridge between the software at the facility and the software that creates the measure. The goal for timing and frequency of data collection is to have data submitted monthly so that performance measures can be made available within three months of when the births occurred.

The data provided by healthcare entities will be provided under a limited data use agreement in order to meet patient confidentiality requirements. The database that stores the records provides encryption of both the date of the birth and the date of the mother´s birth to further improve confidentiality. It is anticipated that the amount of missing data will be significantly decreased through dedicated communication between the healthcare entity and Birthrisk.com, LLC.

3c.2. **Describe any fees, licensing, or other requirements to use any aspect of the measure as specified** *(e.g., value/code set, risk model, programming code, algorithm)*.

A licensing fee would be required for use of the proprietary software. The license fee would depend on how the data is provided. There would be additional costs for creation and maintenance of software bridges. Data on all births would be submitted so that birth data can be accurately reconciled.

The pricing structure is dependent on the method of data entry. If every birth in the United States was submitted on paper then the break-even point is less than one quarter of one percentage point decrease in the cesarean birth rate and if all births were reported via a software bridge then less than one tenth of one percentage point decrease would be the break-even point (assuming a $5,000 savings of a vaginal birth over a cesarean birth). However, if a system of value-based reimbursement is put into place then the break-even point can be achieved immediately. In fact, putting a value-based reimbursement system into place would more than just break-even; it would immediately result in a substantial cost savings.

4. Usability and Use

Extent to which potential audiences (e.g., consumers, purchasers, providers, policy makers) are using or could use performance results for both accountability and performance improvement to achieve the goal of high-quality, efficient healthcare for individuals or populations.

4a. Accountability and Transparency

Performance results are used in at least one accountability application within three years after initial endorsement and are publicly reported within six years after initial endorsement (or the data on performance results are available). If not in use at the time of initial endorsement, then a credible plan for implementation within the specified timeframes is provided.

4.1. **Current and Planned Use**

NQF-endorsed measures are expected to be used in at least one accountability application within 3 years and publicly reported within 6 years of initial endorsement in addition to performance improvement.

Planned	Current Use (for current use provide URL)
Public Reporting Quality Improvement with Benchmarking (external benchmarking to multiple organizations) Quality Improvement (Internal to the specific organization)	

4a.1. **For each CURRENT use, checked above, provide:**

- Name of program and sponsor
- Purpose
- Geographic area and number and percentage of accountable entities and patients included

N/A

4a.2. **If not currently publicly reported OR used in at least one other accountability application (e.g., payment program, certification, licensing) what are the reasons?** (*e.g., Do policies or actions of the developer/steward or accountable entities restrict access to performance results or impede implementation?*)

Currently, public reporting of an accurate cesarean birth measure does not exist. Unfortunately, there has been public reporting of fatally flawed cesarean birth measures in both Canada and the United States which have impeded implementation. Persistence at contacting the Canadian Institute for Health Information and a critical analysis published in the Journal of Obstetrics and Gynaecology Canada has led to the removal of their publicly reported fatally flawed cesarean section indicator from their website. Persistence at contacting the Joint Commission of the leveraging error present in their direct standardization age adjustment technique for core measure PC-02 has led to an email acknowledging that they are aware of the error and that they are planning on removing the direct standardization adjustment from their measure. The LeapFrog Group has published a nationwide list of unadjusted NTSV cesarean birth rates by hospital claiming that this measure uses a "tested, validated measure endorsed by the Joint Commission, National Quality Forum and CMS". Critical analysis reveals that an unadjusted NTSV cesarean birth rate has never been tested or validated and is not the measure that was endorsed by the National Quality Forum. Unfortunately, contacting every member of the board at the Leapfrog Group as well as their obstetrical advisory panel to inform them that their cesarean birth measure is fatally flawed and provides a huge disservice to the women who are giving birth has not resulted in any response from the Leapfrog Group.

Several states have requirements on hospitals to report overall and primary cesarean birth rates but these rates are of no practical use to the public. This problem has been discussed with department of health officials in many states and they are looking for an accurate cesarean birth measure. The Joint Commission will be looking to replace core measure PC-02 with a cesarean birth measure that is free from the errors caused by direct standardization and the Leapfrog Group will eventually realize that their measure is fatally flawed. I believe that the ability of the Birthrisk Cesarean Birth Measure to accurately measure the effects of labor management strategies in any target population of women regardless of the actual cesarean birth rate will result in its adoption at the national, state, facility and provider level within a few years after endorsement.

4a.3. **If not currently publicly reported OR used in at least one other accountability application, provide a credible plan for implementation within the expected timeframes -- any accountability application within 3 years and publicly reported within 6 years of initial endorsement.** (*Credible plan includes the specific program, purpose, intended audience, and timeline for implementing the measure within the specified timeframes. A plan for accountability applications addresses mechanisms for data aggregation and reporting.*)

Upon endorsement of the Birthrisk Cesarean Birth Measure the stakeholders who are concerned with decreasing the number of cesarean births across the nation would be made aware that an accurate cesarean birth measure is available. Both the Joint Commission and the LeapFrog group currently work with hospitals across the country with the goal of obtaining and reporting a cesarean birth measure to the public. Either of these organizations would benefit from collaboration. If neither organization collaborates then State Departments of Health would be contacted in an attempt to establish state wide reporting of the cesarean birth measure. In states where state wide reporting is not available, healthcare systems as well as individual hospitals would be contacted for use of the cesarean birth measure for process improvement or public reporting. Large obstetrical group practices can use this measure for process improvement as well.

This performance measure can be made public through the reporting mechanisms of the entities themselves or directly through the internet to the general public if so desired. Access to the general public could be through any device that can access the internet including mobile devices like smart phones.

4b. Improvement

Progress toward achieving the goal of high-quality, efficient healthcare for individuals or populations is demonstrated. If not in use for performance improvement at the time of initial endorsement, then a credible rationale describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.

4b.1. **Progress on Improvement. (Not required for initial endorsement unless available.)**

Performance results on this measure (current and over time) should be provided in 1b.2 and 1b.4. Discuss:

- **Progress (trends in performance results, number and percentage of people receiving high-quality healthcare)**
- **Geographic area and number and percentage of accountable entities and patients included**

N/A

4b.2. **If no improvement was demonstrated, what are the reasons? If not in use for performance improvement at the time of initial endorsement, provide a credible rationale that describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.**

Improving the quality of care provided to women who labor is ultimately up to the obstetrical care provider. Performance results provided at the provider level reveal that performance scores at the facility level could be misleading. For example, performance measure scores provided for the facility with the most births revealed that almost 90% of the providers at that facility had average performance scores and that the below average score of the facility was a result of the performance of only about 10% of the providers. Having a performance measure that provides provider specific results will allow for quality improvement at the obstetrical care provider level.

4c. Unintended Consequences

The benefits of the performance measure in facilitating progress toward achieving high-quality, efficient healthcare for individuals or populations outweigh evidence of unintended negative consequences to individuals or populations (if such evidence exists).

4c.1. **Were any unintended negative consequences to individuals or populations identified during testing; OR has evidence of unintended negative consequences to individuals or populations been reported since implementation? If so, identify the negative unintended consequences and describe how benefits outweigh them or actions taken to mitigate them.**

N/A

5. Comparison to Related or Competing Measures

If a measure meets the above criteria and there are endorsed or new related measures (either the same measure focus or the same target population) or competing measures (both the same measure focus and the same target population), the measures are compared to address harmonization and/or selection of the best measure.

5. Relation to Other NQF-endorsed Measures

Are there related measures (conceptually, either same measure focus or target population) or competing measures (conceptually both the same measure focus and same target population)? If yes, list the NQF # and title of all related and/or competing measures.

Yes

5.1a. **List of related or competing measures (selected from NQF-endorsed measures)**

0471 : PC-02 Cesarean Birth

5.1b. **If related or competing measures are not NQF endorsed please indicate measure title and steward.**

5a. Harmonization

The measure specifications are harmonized with related measures;

OR

The differences in specifications are justified

5a.1. **If this measure conceptually addresses EITHER the same measure focus OR the same target population as NQF-endorsed measure(s):**

Are the measure specifications completely harmonized?

5a.2. **If the measure specifications are not completely harmonized, identify the differences, rationale, and impact on interpretability and data collection burden.**

5b. Competing Measures

The measure is superior to competing measures (e.g., is a more valid or efficient way to measure);

OR

Multiple measures are justified.

5b.1. **If this measure conceptually addresses both the same measure focus and the same target population as NQF-endorsed measure(s):**

Describe why this measure is superior to competing measures (e.g., a more valid or efficient way to measure quality); OR provide a rationale for the additive value of endorsing an additional measure. (Provide analyses when possible.)

The Birthrisk Cesarean Birth Measure is superior to measure #0471 for several reasons.

The case mix used by the Birthrisk Cesarean Birth Measure includes approximately two thirds of pregnant women whereas the case mix for measure #0471 only includes about one third. Measure #0471 uses a case mix suggested by the American College of Obstetricians and Gynecologists (ACOG) in their Evaluation of Cesarean Delivery from 2000. However, ACOG also stated that "The highest variation occurs among nulliparous patients with term singleton fetuses with vertex presentations (NTSV) without other complications." and "Differences in patient characteristics probably account for some of the variations in cesarean delivery rates and explain some of the differences between practitioners and hospitals." in their Evaluation of Cesarean Delivery. Measure #0471 ignores ACOG's statements concerning patients with "other complications" as well as the significant effect on the risk for cesarean birth due to differences in patient characteristics. The Birthrisk Cesarean Birth Measure includes the NTSV pregnancies that are without other complications and accounts for differences in patient characteristics. Additionally, the Birthrisk Cesarean Birth Measure includes women who have already had a prior vaginal birth. The case mix used by the Birthrisk Cesarean Birth Measure better reflects the statements made by ACOG in 2000 as to NTSV pregnancies and will allow for the improvement of care to not only nulliparous women but also to women who have had a prior vaginal birth.

Including patients with "other complications" in the case mix for measure #0471 results in including women who have contraindications for vaginal birth such as placenta previa, fetal distress prior to labor, medical contraindications for labor, fetal contraindications for labor, women with an un-inducible cervix and women who have requested an elective cesarean birth. An

increase or decrease in women with these diagnoses can significantly affect the outcome of the measure resulting in an inaccurate measure of the effect that the obstetrical care provider has on a woman's risk for a cesarean birth. Even one or two additional cesarean births due to these diagnoses can significantly affect the measure as is illustrated in the most important concern below.

Measure #0471 assumes that all nulliparous women with a term single fetus in the vertex position (NTSV) have the same risk for cesarean birth after adjusting for age. However, in addition to maternal age, there are other physical characteristics of the mother and her baby that have been previously proven to significantly affect the risk for a cesarean birth. These include newborn weight, maternal prepregnancy body mass index, maternal height, gestational age and maternal weight gain. In addition, induction of labor has also been previously proven to significantly increase the risk for a cesarean birth. Failure to provide any risk adjustment for all of these previously proven risk factors will result in a misleading measure for obstetrical care providers. For example, analysis of data from millions of women who attempted labor reveals that inducing labor in a five foot two inch 36 year old nulliparous woman with a starting weight of 175 lbs. who has gained 42 lbs. and is carrying a 4,000 gram baby has a cesarean birth rate of approximately 70%. Whereas a five foot four inch 18 year old nulliparous woman with a starting weight of 115 lbs. who has gained 30 lbs. carrying a 3,500 gram baby who arrives in spontaneous labor has a cesarean birth rate of approximately 7%. This tenfold difference in the rate of cesarean birth due to the physical characteristics of the mother and her baby reveals that using an unadjusted or only age adjusted NTSV cesarean birth rate as a cesarean birth measure may result in simply a measure of the physical characteristics of the woman who are giving birth and not a measure of the effect of the obstetrical care provider's labor management strategies.

The most important concern is a major flaw found in the direct standardization technique being used to create the risk adjustment for age. The direct standardization technique used in measure #0471 is based on the work of Main et al. from 2006. The flaw in the direct standardization technique is illustrated by the sample hospital in their study. The sample hospital in their study had approximately 18,000 births over a three year period in order to create a target population of 7,068 nulliparous term singleton vertex (NTSV) births of which only 68 were in the 15 to 19 year old age group. This age group is assigned a weight of 21% in the direct standardization. This means that even though the sample hospital only had 1% of their births in the 15 to 19 year old age group this age group will be used to assign 21% of their cesarean birth measure. A small change in the number of cesarean births within that age group will result in a large change in their cesarean birth measure. Even with 6,000 total births each year the sample hospital will only have two patients per month and six patients per quarter accounting for 21% of their cesarean birth measure. This will make it very difficult for the sample hospital to obtain consistent results and this problem would only be magnified if the hospital had fewer than 6,000 births per year. If a hospital has the same age distribution of NTSV patients as the sample hospital in the study, critical analysis reveals that one additional cesarean birth in the 15 to 19 year old age group per 1,000 total births will increase their measure #0471 by five percentage points. This flaw makes measure #0471 meaningless not only for hospitals that have an age distribution that is similar to the sample hospital but also for hospitals whose age distribution is not similar to the national average.

Lastly, the goal of a cesarean birth measure is to measure the effect applied by the labor management strategies used by the obstetrical care provider. The accuracy of a cesarean birth measure is best validated by proving its ability to use the measure to predict future outcomes. Despite the fact that measure #0471 was developed many years ago this measure has never been validated by using it to accurately predict future outcomes. In fact, measure #0471 relies only on face validity. A cesarean birth measure that cannot accurately predict future outcomes is merely an educated guess of the risk applied by the obstetrical care provider and not truly a measure of the effect of the obstetrical care provider's labor management strategies.

Appendix

A.1 Supplemental materials may be provided in an appendix. All supplemental materials (such as data collection instrument or methodology reports) should be organized in one file with a table of contents or bookmarks. If material pertains to a specific submission form number, that should be indicated. Requested information should be provided in the submission form and required attachments. There is no guarantee that supplemental materials will be reviewed.
Available at measure-specific web page URL identified in S.1 **Attachment:**

Contact Information

Co.1 **Measure Steward (Intellectual Property Owner):** Birthrisk.com, LLC.
Co.2 **Point of Contact:** Gustavo, San Román, M.D., DoctorGus@Birthrisk.com, 631-331-8885-248
Co.3 **Measure Developer if different from Measure Steward:** Birthrisk.com, LLC.
Co.4 **Point of Contact:** Gustavo, San Román, M.D., DoctorGus@Birthrisk.com, 631-331-8885-248

Additional Information

Ad.1 **Workgroup/Expert Panel involved in measure development** **Provide a list of sponsoring organizations and workgroup/panel members' names and organizations. Describe the members' role in measure development.**
Measure Developer/Steward Updates and Ongoing Maintenance Ad.2 **Year the measure was first released:** Ad.3 **Month and Year of most recent revision:** Ad.4 **What is your frequency for review/update of this measure?** Ad.5 **When is the next scheduled review/update for this measure?**
Ad.6 **Copyright statement:** Copyright 2007 by Birthrisk.com, LLC. All rights reserved. Ad.7 **Disclaimers:**
Ad.8 **Additional Information/Comments:**

MEASURE WORKSHEET

This document summarizes the evaluation of the measure as it progresses through NQF's Consensus Development Process (CDP). The information submitted by measure developers/stewards is included after the Brief Measure Information, Preliminary Analysis, and Pre-meeting Public and Member Comments sections.

To navigate the links in the worksheet: Ctrl + click link to go to the link; ALT + LEFT ARROW to return

Brief Measure Information

NQF #: 2893
Measure Title: Neonatal Intensive Care All-Condition Readmissions
Measure Steward: The Children´s Hospital of Philadelphia
Brief Description of Measure: The NICU Readmissions metric assess the hospital- or state-level readmission rate at 30 days after a stay in the Neonatal Intensive Care Unit.
Developer Rationale: Hospital readmissions have been an area of particular interest for state and national policy agencies, health insurers, and caregivers because of the high costs, both financial and to families, associated with them. So-called "preventable" readmissions, described because ostensibly some change in practice at either the inpatient or outpatient level could have prevented the readmission, may provide insight into care practices that could limit hospitalizations. In pediatric medicine, there are groups of high-risk patients for whom hospital readmissions occur frequently. For example, while the estimated readmission rate within 30 days among the 2.4 million admissions annually in the United States is approximately 6.5% (Berry 2013, Yu 2011), many conditions such as surgery, sickle cell disease, and prematurity have rates between 15 and 20% (Ray 2013, Wade 2008, Underwood 2007, Berry 2014). One other group is children discharged from the neonatal intensive care unit (NICU). Premature infants have an approximately three-fold increase in risk of hospital readmission after discharge compared to term infants, with higher rates in infants of younger gestational age (Ray 2013). These hospital readmissions contribute to the higher health care costs and utilization seen in prematurely-born infants (Wade 2008, Underwood 2007). For the preterm population, we have demonstrated substantial variation in the unadjusted readmission rates among California hospitals: Mean rate of 5.8%; median of 4.5%; Interquartile range 2.5%-7.8%; Range 0%-50%; Standardized difference 909% for the range, and 97% for the interquartile range. The large variation between hospitals persisted after adjusting for gestational age and sociodemographic factors: mean rate 6.0%; median 4.8%; interquartile range 2.5%-7.9%; range 0%-52.8%; standardized difference 943% for the range, and 96% for the interquartile range. Of note, the average percent change between the unadjusted and risk-adjusted rates was 0.75% (that is, the risk-adjusted rate was less than 1% different from the unadjusted rate; for the median unadjusted rate of 4.5%, this would result in a risk-adjusted rate of 4.545%), with a range of -3.4% to 5.6%.
Similar differences are seen between states within their Medicaid population, suggesting that larger systems of care beyond the NICU itself may be important to optimize the health of infants who require neonatal intensive care. These systems may be at the health system level or at the level of the outpatient care provider (Lorch 2010).
Based on several conceptual frameworks for readmissions in the pediatric or specific neonatal population, readmission rates can be considered a product of:
- patient factors, such as illness severity at the time of discharge and socioeconomic factors at the patient and community level, such as access to care and health insurance; and
- inpatient quality of care provided during the hospitalization, which may impact readmissions either directly, or more likely through altering the severity of illness of a specific patient at the time of discharge through differing rates of bronchopulmonary dysplasia, necrotizing entercolitis, or intraventricular hemorrhage, each of which has been associated with higher readmission rates in prematurely-born infants (Lorch 2010, Ray 2012, Lorch 2014); and
- care processes around the time of discharge, including education of families to prepare them for discharge and processes to ensure the transition of care from the inpatient to the outpatient setting; and
- outpatient care quality (Lorch 2010).

Therefore, implementing a quality measure around readmission rates from the neonatal intensive care unit will result in improved care around the transition from inpatient to outpatient care; strengthen the post-discharge support for the infants who required neonatal intensive care; and decrease overall costs of health care association with variation in rehospitalization rates of these infants, especially if the readmission measure assesses a different aspect of quality from other potential measures, such as complications of preterm birth. This assessment will be demonstrated with poor correlation between readmission rates and hospital complication rates, both risk adjusted and unadjusted.

Numerator Statement: Number of infants with a gestational age between 23-34 weeks who were readmitted to the hospital within 30 days of discharge. These time periods are assessed cumulatively, such that readmissions occurring within prior time periods are included. Reliability is strongest if each health care unit has at least 50 discharges per time unit studied.
Denominator Statement: Number of newborns with a gestational age between 23-34 weeks discharged from the NICU, based on gestational age field contained in the birth certificate record (best obstetrical estimate).
Denominator Exclusions: Infants with a specified congenital anomaly are excluded from the target population.
Infants with a missing gestational age are excluded from the primary analysis. Information about multiple imputation methods to allow for their inclusion are presented in the testing attachment, section 2b7.
Infants who expired during the neonatal intensive care period are not eligible for a hospital readmission and excluded.

Measure Type: Outcome
Data Source: Administrative claims, Electronic Clinical Data : Electronic Health Record, Other
Level of Analysis: Facility, Population : State

IF Endorsement Maintenance – Original Endorsement Date: Most Recent Endorsement Date:

New Measure -- Preliminary Analysis

Criteria 1: Importance to Measure and Report

1a. Evidence

1a. Evidence. The evidence requirements for a health outcomes measure include providing rationale that supports the relationship of the health outcome to processes or structures of care. The guidance for evaluating the clinical evidence asks if the relationship between the measured health outcome and at least one clinical action is identified and supported by the stated rationale.

Summary of evidence: The developers provide a conceptual logic model and cite three studies that support their statement "On the hospital level, readmissions represent either poor quality of care during the hospitalization, or poor discharge planning and transition of care from inpatient to outpatient providers."

Question for the Committee:

- *Is there at least one thing that the provider can do to achieve a change in the measure results?*

Preliminary rating for evidence: ☒ **Pass** ☐ **No Pass**

1b. Gap in Care/Opportunity for Improvement and 1b. Disparities
Maintenance measures – increased emphasis on gap and variation

1b. Performance Gap. The performance gap requirements include demonstrating quality problems and opportunity for improvement.

- The developers provide descriptions of the hospitals providing data for 1995 – 2009 in Californiain the Appendix Table 1.
- The measure results for 30 day readmission with descriptive statistics in the Appendix, Table 3.

A sample of data from the table shows the most recent aggregate measure results:

	Mean	Std Dev	Min	Max	IQR
2006	0.099	0.101	0.000	0.865	(0.055-0.12)
2007	0.092	0.074	0.000	0.501	(0.049-0.119)
2008	0.102	0.107	0.000	1.212	(0.046-0.126)
2009	0.093	0.099	0.000	1.125	(0.044-0.126)

The mean result in 2009 shows that 9.3% were readmitted. The value of 1.125 for the 2009 Max is 112%. According to the developer " *The method of calculating the risk adjusted rate that was also pulled does occasionally result in a rate above 100% if the observed/expected rate of readmissions is substantially above 1. Thus, this high rate suggests a very large observed rate of readmissions that was not expected based on clinical data.* **"**

Disparities

The developers provide measure results stratified by race/ethnicity (CL = confidence limits). A sample is provided here:

		Readmission 30 Days Rate (CL)
2007	White	0.075 (0.068-0.081)
	Black	0.112 (0.098-0.127)
	Hispanic	0.096 (0.09-0.102)
	Other	0.072 (0.063-0.082)
2008	White	0.077 (0.071-0.085)
	Black	0.102 (0.088-0.117)
	Hispanic	0.096 (0.09-0.102)
	Other	0.075 (0.065-0.085)
2009	White	0.078 (0.071-0.086)
	Black	0.111 (0.097-0.128)
	Hispanic	0.099 (0.093-0.105)
	Other	0.066 (0.057-0.076)

Questions for the Committee:

- *Is there a gap in care that warrants a national performance measure?*
- *Does this measure provide useful information about disparities in peritnatal care?*

Preliminary rating for opportunity for improvement: ☒ **High** ☐ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 1: Importance to Measure and Report (including 1a, 1b, 1c)

1a. Evidence to Support Measure Focus

Comments:

The relationship between hospital readmission and processes like poor discharge planning and transition of care have been identified by developers and well-supported.

The developer hypothesizes a lower NICU-grad 30-day risk-adjusted readmission rate is associated with better care.

1b. Performance Gap

Comments:

Yes, significant performance gap and disparities noted in the descriptions provided.

**NICU-grads have a higher risk of preventable readmission than term neonates: illness severity, inpatient care, socioeconomic status/insurance status, outpatient provider quality, and transition-to-home care may all impact

readmission rates. However, the readmission rates vary very widely among the institutions studied (CA, UT, NY) It is not clear that higher readmission rates equates to poorer neonatal care.**

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability

2a1. Reliability Specifications

2a1. Specifications requires the measure, as specified, to produce consistent (reliable) and credible (valid) results about the quality of care when implemented.

Data source(s): Administrative claims, Electronic Clinical Data : Electronic Health Record, Other

Specifications:

- The level of analysis is facility or state.
- The measure result is a ratio:
  - The observed number of readmissions divided by the expected number of readmissions at the given level of the health care provider (hospital, outpatient facility, state) where the expected number of readmissions is based on the risk adjustment model.
  - The final ratio, (observed number of readmissions)/(expected number of readmissions), is then multiplied by the mean percentage of admissions in the cohort to calculate a risk-adjusted readmission rate.
- Three exclusions:
  - Infants with a specified congenital anomaly are excluded from the target population.
  - Infants with a missing gestational age are excluded from the primary analysis. Information about multiple imputation methods to allow for their inclusion are presented in the testing attachment, section 2b7.
  - Infants who expired during the neonatal intensive care period are not eligible for a hospital readmission and excluded
- A data dictionary is include that contains codes for NICU type, complications_comorbidities and congenital anomalies.
- A calculation algorithm is decribed.
- The measure is risk-adjusted for gestational age, infant gender, maternal race/ethnicity, complilcations of prematurity (BPD, ROP, NEC, IVH), maternal age and source of payment.

Questions for the Committee :

- *Are all the data elements clearly defined? Are all appropriate codes included?*
- *Are codes needed to identify the denominator "discharged" from the NICU?*
- *How are readmissions captured?*
- *Is the logic or calculation algorithm clear?*
- *Is it likely this measure can be consistently implemented?*

2a2. Reliability Testing Testing attachment

2a2. Reliability testing demonstrates if the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise enough to distinguish differences in performance across providers.

SUMMARY OF TESTING

Reliability testing level ☒ **Measure score** ☐ **Data element** ☐ **Both**

Reliability testing performed with the data source and level of analysis indicated for this measure ☐ **Yes** ☐ **No**

Method(s) of reliability testing

- The measure was tested using hospital data from the California Patient Discharge Data/Emergency Department/Ambulatory Surgery/Birth Cohort linked with vital statistics data.
- Description of the test population is provided. In 2009 there were 252 hospitals and 18,583 patients.
- ***Reproducibility*** of the results were calculated using Spearman-Brown statistics. Briefly, a 50% random sample of patients was drawn from each health care unit (hospital), and risk-adjusted rates were calculated. Then, a second 50% sample was chosen and Spearman rank sum correlation coefficients were calculated. This is an appropriate method for testing reliability.

Results of reliability testing

- Spearman-Brown correlation coefficients are presented by hospital volume quartile because some of the smallest hospitals have few events to examine.

Readmission 30 Day	Spearman-Brown correlation
Volume Lowest Quartile	0.736
Volume 2nd Quartile	0.931
Volume 3rd Quartile	0.880
Volume Highest Quartile	0.963

- The developers conclude that ‘‘Data show acceptable reliability, with Spearman-Brown split sample correlations above 0.70 for each volume strata. Stronger reliabilities above 0.85 are found in the three largest volume strata.” This is an appropriate test for reliability. Generally, reliability > 0.70 is considered acceptable.
- In the numerator statement the developer states “Reliability is strongest if each health care unit has at least 50 discharges per time unit studied.”

Guidance from the Reliability Algorithm

Precise specifications -some questions about all necessary codes (Box 1)→ empirical testing with statistics (Box 2) →testing of computed measure score (Box 4) → appropriate method (Box 5) → high (Box 6a)

Questions for the Committee:

- *Is the test sample adequate to generalize for widespread implementation?*
- *Do the results demonstrate sufficient reliability so that differences in performance can be identified?*

Preliminary rating for reliability: ☒ **High** ☐ **Moderate** ☐ **Low** ☐ **Insufficient**

2b. Validity

2b1. Validity: Specifications

2b1. Validity Specifications. This section should determine if the measure specifications are consistent with the evidence.

Specifications consistent with evidence in 1a. ☒ **Yes** ☐ **Somewhat** ☐ **No**

Question for the Committee:

- *Are the specifications consistent with the evidence?*

2b2. Validity testing

2b2. Validity Testing should demonstrate the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality.

SUMMARY OF TESTING

Validity testing level ☒ **Measure score** ☐ **Data element testing against a gold standard** ☐ **Both**

Method of validity testing of the measure score:

- ☐ **Face validity only**
- ☒ **Empirical validity testing of the measure score**

Validity testing method:

- Two hypotheses were tested to establish the validity of this measure:
  - ***Hypothesis 1: there should be a larger association between hospital volume and 30-day readmission rates compared to complication rates. This work parallels other work in the literature that suggests that higher volume hospitals have improved outcomes, likely secondarily to seeing more patients and implementing processes of care to improve their outcomes***.
    - Using the California hospital data set described above, correlation between 30-day readmission rates and hospital volume was explored.
  - ***Hypothesis 2: The processes of care that may reduce the development of complications (BPD, ROP, NEC, IVH) throughout the hospital stay (improved respiratory ventilator management, improved feeding programs, hand hygiene programs) may only be modestly associated with processes of care that improve readmission rates (such as improved transitions of care, improved education of families, etc).***
    - Using the California hospital data set described above, correlation between 30-day readmission rates and complications of prematurity was explored.

Validity testing results:

- Correlation with hospital volume:

Table 3. Hospital Volume Correlation	
Quintile	30-day risk adjusted readmission rate
Lowest	0.062
2	0.057
3	0.053
4	0.051
Highest	0.052
P-value, test of trend	0.032

  - With a greater number of hospitals, we found a volume-outcome association with rates of readmission within 30 days after discharge: lower volume hospitals had higher rates of readmissions through this time period. While the absolute difference was 1.0% between the smallest and largest quintile, this difference was a 16% relative increase in readmission rates. Hypothesis generally supported.

- Correlation with complications of prematurity:

Table 4. Correlation of Adjusted Readmission Rate with Complication Rates, Spearman correlation coefficients	
BPD	0.12
ROP	0.14
NEC	0.12
IVH	0.16

All correlations had a p-value > 0.05.

- There were poor to no correlations between hospital-level risk adjusted rates of complications and risk-adjusted readmission rates. Such lack of correlations is similar to that found in most studies of adult readmission rates, which hypothesize that readmission rates assess a different aspect of the quality of NICUs compared with that assessed by complication rates. Confirms hypoethsis of no correlation.

Questions for the Committee:

- *Is the test sample adequate to generalize for widespread implementation?*
- *Do the results demonstrate sufficient validity so that conclusions about quality can be made?*
- *Do you agree that the score from this measure as specified is an indicator of quality?*

2b3-2b7. Threats to Validity

2b3. Exclusions:

- The developers provide data on frequency of exlcusions except for missing gestational age data:

	Patients Excluded for Congenital Anomalies	Hospitals Excluded for small volume, 1 year
2006	1,598 (6.84%)	0 (0.0%)
2007	1,665 (7.14%)	5 (0.02%)
2008	1,610 (7.46%)	1 (0%)
2009	1,611 (7.98%)	2 (0.01%)

- The developers report that the rates of readmissions were substantively higher in the congenital anomaly group, with a mean increase of 9% higher.

Questions for the Committee:

- *Are the exclusions consistent with the evidence?*
- *Are any patients or patient groups inappropriately excluded from the measure?*
- *Are the exclusions/exceptions of sufficient frequency and variation across providers to be needed (and outweigh the data collection burden)?*

2b4. Risk adjustment: **Risk-adjustment method** ☐ **None** ☒ **Statistical model** ☐ **Stratification**

Conceptual rationale for SDS factors included ? ☒ **Yes** ☐ **No**

- The developers note "Sociodemographic data were included because of prior data from our group and others (Lorch Academic Pediatrics 2014; Ray Hospital Pediatrics 2013; Ray Academic Pediatrics 2010) that have found an association between readmission and maternal age, likely secondary to differences in provider practice related to a provider's assessment of the compliance of the family and the parental preferences toward a hospital admission; and higher rates of readmission in publicly-insured children. We included race/ethnicity in the models for completeness sake and because there were identified differences in these data."

SDS factors included in risk model? ☒ **Yes** ☐ **No**

- SDS factors were included based on their prevalence in the dataset and their association with the outcome as noted in the risk factor table.
- Maternal race/ethnicity and source of payment were included as risk-factors.

Risk adjustment summary

- Risk factors in clude gestational age, complications of prematurity, maternal race/ethnicity, maternal age and source of payment.
- The c-statistic result for 30-day readmission was found to be 0.58. The c-statistic is model discrimination statistic that represents the proportion of all-possible pairs with different observed outcomes for which the model correctly predicts a higher probability for observations with the outcome of interest than those without the outcome of interest. A *c*-statistic of 0.58 means that for 58% of all possible pairs of patients—one who dwas readmitted and one who was not re-admitted—the model correctly assigned a higher probability to those who were readmitted. Generally, a c-statistic of at least 0.70 is considered acceptable. .
- The developers note "The risk adjustment model shows modest discrimination and strong calibration. While the c-statistic for the risk adjustment model is less than the desired level of 0.7, this model for predicting readmission risk using medical factors is similar to prior work from our group and others, with a c-statistic between 0.6 and 0.65, and in other outcome measures such as length of stay and health care costs (Ray KN. Hospital Pediatrics. 2013;3:194-203; Lorch SA. Hlth Serv Res. 2010;45:24-41.; Schwartz, M., and A. S. Ash. 2003. ''Evaluating Risk-Adjustment Model Empirically.'' In Risk Adjustment for Measuring Health Care Outcomes, edited by L. I. Iezzoni. Chicago: Health Administration Press)."
- The developers also explain that they "initially considered a non-risk adjusted metric. This unadjusted rate argues that any readmission is undesired – although a zero percent readmission rate is not practical to achieve and may have other untoward consequences such as longer initial lengths of stay. However, to improve face validity, we included factors that have been associated with higher readmission rates in the literature. Of note, the average percent change between the unadjusted and risk-adjusted rates was 0.75%, with a range of -3.4% to 5.6%. Thus, the inclusion of risk adjustment was felt to be necessary by clinicians to improve the face validity of the results, but made little difference to the actual measurements.

Questions for the Committee:

- *Is an appropriate risk-adjustment strategy included in the measure?*
- *Are the candidate and final variables included in the risk adjustment model adequately described for the measure to be implemented?*
- *Are all of the risk adjustment variables present at the start of care? If not, describe the rationale provided.*

2b5. Meaningful difference *(can statistically significant and clinically/practically meaningful differences in performance measure scores can be identified)*:

- The developers indicate that "Outliers were defined as those entities with rates greater than 75th percentile value + 1.5*interquartile range (aka "upper adjacent value"), or less than 25th percentile - 1.5*interquartile range (aka "lower adjacent value")."
- The developer reports " All outlier hospitals had rates above the upper adjusted rate. In each case, approximately 8% of the hospital population met the definition described above."

Question for the Committee:

- *Does this measure identify meaningful differences about quality?*
- *Is 8% outliers sufficient for this measure to identify meaningful differences in performance?*

2b6. Comparability of data sources/methods: Developer indicates NA.

2b7. Missing Data

- After testing imputed data for gestational age and finding a high correlation of measure results with missing gestation age excluded vs imputed, the developers have chosen to exclude missing values for gestational age. Cases with a missing gestational age had lower readmission rates "likely because patients missing GA tended to be of older gestational age using multiple imputation techniques".
- Frequency of missing gestational age was not provided.

Guidance from the algorithm:
Consistent with specifications (Box 1) →potential threats to validity addressed (Box 2) → empirical testing (Box 3) →testing of meaure score (Box 6) →appropriate testing method (Box 7) → testing results moderate → moderate

Preliminary rating for validity: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 2: Scientific Acceptability of Measure Properties (including all 2a, 2b, and 2d)

2a1. & 2b1. Specifications

Comments:

The measure title "Neonatal Intensive Care All-Condition Readmissions" is misleading as only infants born between gestational age 23-34 weeks are captured. Gestational age may not always be captured in the neonate's chart or captured accurately. Moreover, often there are discrepancies between the obstetric estimate vs. the gestational age assigned after physical exam and Ballard scoring in the NICU. Would it be better to capture all NICU admissions rather than restrict it to 23-34 weeks gestation?

Elements define. Exclusion diagnoses not provided.

Adult CHF 30 day readmission reductions target post-discharge care coordination. Recent data suggest that highly-integrated discharge planning (pharmacy/nursing/PCP/home health/follow up phone calls) resulted in a paradoxical increase in readmission rates in at least one study. (Bradley EH, et al. Circ Cardiovasc Qual Outcomes. 2013 July; 6(4):444-450.) Would similar intensive discharge planning (costly) drive up post-NICU readmissions (more costly) unintentionally? The prevalence of CHF readmits (national priority) versus NICU-grads reveal a much smaller scale of recoverable costs by reducing neonatal readmits.

2a2. Reliability Testing

Comments:

Spearman correlation was performed. However, this reviewer did not appreciate any normative values defined by the developer.

2b2. Validity Testing

Comments:

C-statistical analysis revealed less-than-desired 0.7. This reviewer is not convinced that the score, as measured, correlates with quality of care provided in the NICU setting. From the measure developer's own body of work, he has demonstrated that the post-hospital/outpatient setting drive readmission rates to a larger degree than the NICU can. So why hold a NICU/institution accountable for something it has little ability to influence?

**I cannot support this measure development as currently presented: Validity is Low to moderate.

The value to compare institutional readmission rates that may/may not be related to the quality of NICU care (but more likely to discharge planning, patient-specific risk-adjusted health conditions, availability of outpatient resources) is not compelling based on the data presented.**

2b3. Exclusions Analysis

2b4. Risk Adjustment/Stratification for Outcome or Resource Use Measures

2b5. Identification of Statistically Significant & Meaningful Differences In Performance

2b6. Comparability of Performance Scores When More Than One Set of Specifications

2b7. Missing Data Analysis and Minimizing Bias

Comments:

**Threats to validity: Many maternity hospitals with large NICUs do not have a pediatric unit and hence these infants will not be re-admitted to the same hospital. Capturing this information from multiple hospitals can be challenging especially when the last names are changed from mother's maiden name to father's last name, insurance company switching and movement of families from one geographic location to another.

Risk adjustment: Birth weight, sex of the infant and use of antenatal steroids are well-recognized risk factors that affect neonatal mortality and morbidity. These factors have not been included in the risk-adjustment model.**
**There was no set of ICD10 codes for congenital exclusions provided to this reviewer.
There was no estimate for the number of patients excluded, But I cannot comment on the statistical analysis for this excluded data. pg 40.**

Criterion 3. Feasibility

3. Feasibility is the extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

- The developer reports that "that implementation may be difficult due to missing data from administrative datasets in use at the state and federal level. Important adjusting variables such as gestational age and birth weight are not currently recorded consistently in MAX or like datasets, and thus accurate implementation of this metric will require new data collection, linkage with birth certificates, or more widespread and standardized use of EHR for publicly reported measures."

Questions for the Committee:

- *Is linking administrative data to birth records available outside of California?*
- *Are the required data elements routinely generated and used during care delivery?*
- *Are the required data elements available in electronic form, e.g., EHR or other electronic sources?*
- *Is the data collection strategy ready to be put into operational use?*

Preliminary rating for feasibility: ☐ **High** ☐ **Moderate** ☒ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 3: Feasibility

3a. Byproduct of Care Processes
3b. Electronic Sources
3c. Data Collection Strategy
Comments:
Same as in 2b 3-7 response.
This measure uses data set that must be newly reported; not existing data set. Decreases feasibility.

Criterion 4: Usability and Use

4. Usability and Use evaluate the extent to which audiences (e.g., consumers, purchasers, providers, policymakers) use or could use performance results for both accountability and performance improvement activities.

Current uses of the measure [from OPUS]
Publicly reported? ☐ **Yes** ☐ **No**

Current use in an accountability program? ☐ **Yes** ☐ **No**
OR
Planned use in an accountability program? ☒ **Yes** ☐ **No**

Accountability program details

- The developer notes "This metric was developed in part with the CHIPRA Pediatric Quality Measure Program

funded by AHRQ. It will be part of a collection of measures publicized as a resource for policymakers, advocates, consumers, and purchasers of children's health care; researchers; providers; and others interested in identifying valid, reliable, and feasible quality measures of children's health care."
- This metric will be included in the portfolio of measures publicized by the CHIPRA PQMP program, available to all for use.

Improvement results NA

Unexpected findings (positive or negative) during implementation feasibility of collecting data

Potential harms

Feedback :

***Questions for the Committee**:*
- ○ *How can the performance results be used to further the goal of high-quality, efficient healthcare?*
- ○ *Do the benefits of the measure outweigh any potential unintended consequences?*

Preliminary rating for usability and use: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 4: Usability and Use

4a. Accountability and Transparency
4b. Improvement
4c. Unintended Consequences
Comments:
Unintended consequences: Secondary/tertiary care centers providing care for rural/resource-poor primary care setting may be unfairly penalized if measure is tied to reimbursement rates.

Criterion 5: Related and Competing Measures

Related or competing measures

- 2393 : Pediatric All-Condition Readmission Measure

Harmonization

Pre-meeting public and member comments

-

NATIONAL QUALITY FORUM—Evidence (subcriterion 1a)

Measure Number (*if previously endorsed*)**:** Click here to enter NQF number
Measure Title: Neonatal Intensive Care All-Condition Readmissions
IF the measure is a component in a composite performance measure, provide the title of the Composite Measure here: Click here to enter composite measure #/ title

Date of Submission: 3/29/2016

Instructions

- *For composite performance measures:*
  - *A separate evidence form is required for each component measure unless several components were studied together.*
  - *If a component measure is submitted as an individual performance measure, attach the evidence form to the individual measure submission.*
- Respond to all questions as instructed with answers immediately following the question. All information needed to demonstrate meeting the evidence subcriterion (1a) must be in this form. An appendix of *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 10 pages (*incudes questions/instructions*; minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.

Note: The information provided in this form is intended to aid the Steering Committee and other stakeholders in understanding to what degree the evidence for this measure meets NQF's evaluation criteria.

1a. Evidence to Support the Measure Focus
The measure focus is evidence-based, demonstrated as follows:

- Health outcome: [3] a rationale supports the relationship of the health outcome to processes or structures of care. Applies to patient-reported outcomes (PRO), including health-related quality of life/functional status, symptom/symptom burden, experience with care, health-related behavior.
- Intermediate clinical outcome: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured intermediate clinical outcome leads to a desired health outcome.
- Process: [5] a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured process leads to a desired health outcome.
- Structure: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured structure leads to a desired health outcome.
- Efficiency: [6] evidence not required for the resource use component.

Notes

3. Generally, rare event outcomes do not provide adequate information for improvement or discrimination; however, serious reportable events that are compared to zero are appropriate outcomes for public reporting and quality improvement.
4. The preferred systems for grading the evidence are the U.S. Preventive Services Task Force (USPSTF) grading definitions and methods, or Grading of Recommendations, Assessment, Development and Evaluation (GRADE) guidelines.
5. Clinical care processes typically include multiple steps: assess → identify problem/potential problem → choose/plan intervention (with patient input) → provide intervention → evaluate impact on health status. If the measure focus is one step in such a multistep process, the step with the strongest evidence for the link to the desired outcome should be selected as the focus of measurement. Note: A measure focused only on collecting PROM data is not a PRO-PM.
6. Measures of efficiency combine the concepts of resource use and quality (see NQF's Measurement Framework: Evaluating Efficiency Across Episodes of Care; AQA Principles of Efficiency Measures).

1a.1.This is a measure of: (*should be consistent with type of measure entered in De.1*)
Outcome

☒ Health outcome: Readmission
☐Patient-reported outcome (PRO): Click here to name the PRO

PROs include HRQoL/functional status, symptom/symptom burden, experience with care, health-related behaviors

☐ Intermediate clinical outcome (*e.g., lab value*): Click here to name the intermediate outcome

☐ Process: Click here to name the process

☐ Structure: Click here to name the structure

☐ Other: Click here to name what is being measured

_____

HEALTH OUTCOME/PRO PERFORMANCE MEASURE *If not a health outcome or PRO, skip to 1a.3*

1a.2. Briefly state or diagram the path between the health outcome (or PRO) and the healthcare structures, processes, interventions, or services that influence it.

Nakamura et al. (2014) published a conceptual framework diagramming the multifaceted and complex factors that contribute to readmissions (Figure 1). This model illustrated that while medical risk factors, socioeconomic/psychosocial factors, and treatment factors are part of the readmission risk pathway, healthcare structures, processes and services influence rates in tandem. On the hospital level, readmissions represent either poor quality of care during the hospitalization, or poor discharge planning and transition of care from inpatient to outpatient providers (Tsai 2013, Berry 2013, Morse 2011). Data from adult studies suggest that readmissions may better reflect hospital-level variations in the latter construct (Krumholz 2013). Factors within health systems such as quality of inpatient care, timing of discharge, discharge preparation and communication, and outpatient care and services are aspects of care that can be directly addressed (Nakamura 2014, Lorch 2010). While some readmissions are unavoidable, others are possibly predictable and/or entirely preventable. This conceptual model is the framework for the conditionally endorsed Pediatric All-Cause Readmission Measure and serves as a basis for this project.

Figure 1: Nakamura Conceptual Model

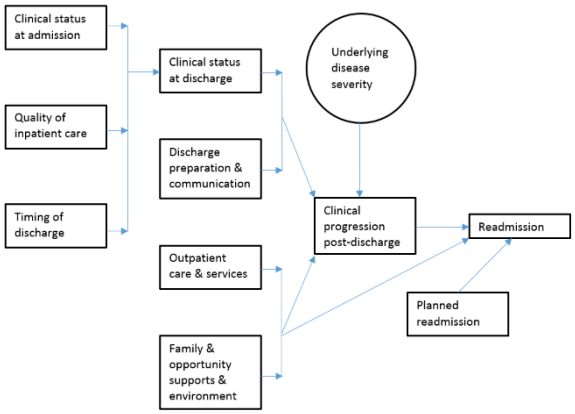
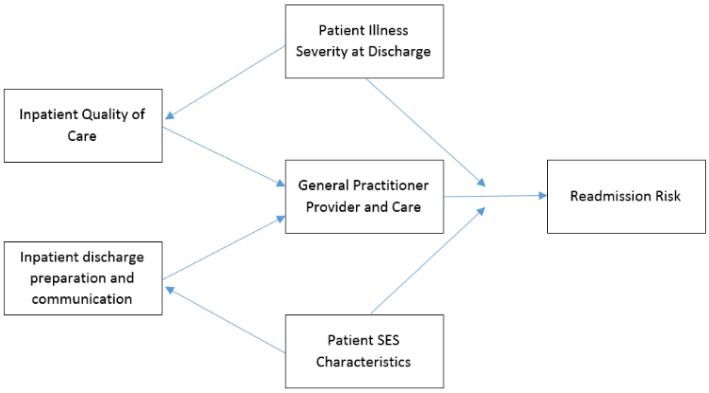


Figure 2: NICU Specific Conceptual Model



A conceptual model specific to the neonatal population, and based on the available evidence both published and tat included in the testing attachment, is shown in Figure 2. Readmission rates can be considered a product of:

- patient factors, such as illness severity at the time of discharge and socioeconomic factors at the patient and community level, such as access to care and health insurance; and
- inpatient quality of care provided during the hospitalization, which may impact readmissions either directly, or more likely through altering the severity of illness of a specific patient at the time of discharge through differing rates of bronchopulmonary dysplasia, necrotizing entercolitis, or intraventricular hemorrhage, each of which has been associated with higher readmission rates in prematurely-born infants (Lorch 2010, Ray 2012, Lorch 2014); and
- care processes around the time of discharge, including education of families to prepare them for discharge and processes to ensure the transition of care from the inpatient to the outpatient setting; and
- outpatient care quality (Lorch 2010).

1a.2.1. State the rationale supporting the relationship between the health outcome (or PRO) to at least one healthcare structure, process, intervention, or service (*i.e., influence on outcome/PRO*).

Hospital readmissions have been an area of particular interest for state and national policy agencies, health insurers, and caregivers because of the high costs, both financial and to families, associated with them. So-called "preventable" readmissions, described because ostensibly some change in practice at either the inpatient or outpatient level could have prevented the readmission, may provide insight into care practices that could limit hospitalizations. In pediatric medicine, there are groups of high-risk patients for whom hospital readmissions occur frequently. For example, while the estimated readmission rate within 30 days among the 2.4 million admissions annually in the United States is approximately 6.5% (Berry 2013, Yu 2011), many conditions such as surgery, sickle cell disease, and prematurity have rates between 15 and 20% (Ray 2013, Wade 2008, Underwood 2007, Berry 2014). One other group is children discharged from the neonatal intensive care unit (NICU). Premature infants have an approximately three-fold increase in risk of hospital readmission after discharge compared to term infants, with higher rates in infants of younger gestational age (Ray 2013). These hospital readmissions contribute to the higher health care costs and utilization seen in prematurely-born infants (Wade 2008, Underwood 2007). A limited number of studies show variation in readmission rates in one Canadian province (Martens 2004) and a small number of hospitals (Morris 2005). Variations in other pediatric specialties has also been shown (Berry 2013, Czaja 2013). For the preterm population, we have demonstrated substantial variation in the unadjusted readmission rates among California hospitals regardless of time period examined with a standardized difference that ranged from 578%-683%. The large variation between hospitals persisted after adjusting for gestational age and sociodemographic factors, with standardized differences again ranging between 660 and 724% (Lorch 2014).

There are a number of potential factors associated with higher, or lower, rates of hospital readmissions (Figure 2). First, differences in rates may result from differences in illness severity (Lorch 2010) or other patient characteristics across hospitals (Lorch 2010, Ambalavanan 2011, Ray 2010). Readmission rates of preterm infants are approximately 3 to 6-fold higher compared to term infants, with the highest rates found in infants of younger gestational age (Ray 2013). Data from numerous adult studies show that infants of lower socioeconomic status have higher rates of hospital readmissions (Srivastava 2013), leading to higher rates of readmissions at safety-net hospitals for surgical procedures (Hoehn 2015) and congestive heart failure (Joynt 2010). Other studies show associations between readmission rates within ZIP codes and rates of poverty and other measures of social deprivation (Beck 2012, Ray 2012, Liu 2009). Family socioeconomic status, as measured by insurance status (Rice-Townsend 2013, Liu 2009, Bloomberg 2003, Auger 2013, Coller 2013) and financial hardship (McGregor 2006), is associated with readmission risk. Children with publicly-financed insurance have higher rates of readmission, with prematurely-born infants in some states having rates as high as 30% (Lorch 2014). For this reason, it is important that the readmission metric be risk-adjusted for all health care groupings smaller than the state level (see testing attachment for information about the reliability of both unadjusted and risk-adjusted metrics).

Readmissions may also result from differences in outpatient providers and practices after discharge. For example, recent data from our group found increased rates of hospital readmission for preterm infants receiving care at outpatient providers with a higher use of unnecessary antibiotics or other medications (Lorch 2010). Coller, in a systematic review of pediatric hospitals published in 2014, found that the primary method of preventing readmissions in children with complex health issues was improved continuity and care coordination (Coller 2014). When studies account

for all aspects of the health care system, both inpatient and outpatient, it may inform observed inter-hospital differences in readmission rates shown by our data (Figures in Section 2b5.2. of the testing attachment) and others using Medicare (Herrin 2015) or individual hospital data (McMillan 2015).

Studies exploring the association of hospital readmission rates and overall measures of quality at the hospital level have found conflicting, and sometimes surprising results. Many studies have failed to show an association between hospital complication rates and readmission rates (Yeh 2012, Li 2012, Brown 2014, Horwitz 2015). We show similar results for rates of bronchopulmonary dysplasia, necrotizing enterocolitis, and retinopathy of prematurity (Tables in Section 2b2.3 of the Testing Form). Prior work has also suggested that higher quality NICUs have higher volumes and lower complication rates than their peers (Phibbs 1997, Phibbs 2007). However, for readmissions, hospital volume has shown conflicting results, with some studies suggesting that high volume hospitals have lower rates of readmission (Brown 2014, Tsai 2013), and other studies finding the opposite association (Joynt 2011, Horwitz 2015). However, as we and others have argued (Lorch 2013, Krumholz 2013) this lack of associations with some measures of hospital quality may reflect the theory that readmission measures a different aspect of care – the discharge/transition to home process, including education of the family – that fails to affect other quality measures such as complication or mortality rates (Coller 2014).

Key factors that may affect the structures or processes of care that the readmission metric assesses are (a) the cohort of patients included in the study and (b) the time frame for readmission after hospital discharge. Compared to other readmission measures, including only infants admitted to the NICU allows us to reduce the noise within the measure because of differences in admission criteria between hospitals. All infants born under 34 weeks of gestational age are included when gestational age is in the data source or can be imputed. Ideally, gestational age is known when determining sample inclusion. CMS uses a 30-day time frame for their readmission measures (CMS 2013). However, there are no studies to support this time frame over shorter (Escobar 1999) or longer evaluation periods used in prior work (Ray 2013, Lorch 2010). Shorter time periods may better reflect the care delivered by the inpatient hospital course, suggested by our reliability data shown in the testing attachment, whereas longer time periods may reflect either care of outpatient providers or the overall illness severity of these infants (Lorch 2010). Thus, we define the metric with rehospitalizations for any reason within 7, 14, 30, 90, and 365 days after discharge from the birth hospitalization, depending on the goals of the group implementing the metric.

Overall, then there have been no comprehensive studies of other structural metrics or processes of care associated with differences in readmission rates, particularly at the level of the neonatal intensive care unit. However, the observed substantial hospital and state-level variation seen in this work (Lorch 2014), which occurs after adjusting for patient-level factors such as sociodemographic factors and clinical variables, supports the idea that readmission variation reflects some part of a hospital's ability to transition care from the inpatient to outpatient setting, along with the outpatient provider's ability to accept and manage the patient. Specific areas that these measures may assess include (1) specific transition of care policies and protocols; (2) education of families; (3) choice of outpatient provider; (4) communication between inpatient and outpatient provider; (5) access to outpatient care (Misky 2010); or (6) quality of outpatient provider in reducing readmission risk.

References

1. Ambalavanan N, Carlo WA, McDonald SA, Yao Q, Das A, Higgins RD, Generic D, Follow-up Subcommittees of the Eunice Kennedy Shriver National Institute of Child H, Human Development Neonatal Research N. Identification of extremely premature infants at high risk of rehospitalization. Pediatrics 2011; 128:e1216-1225. PMC3208965.
2. Auger KA, Kahn RS, Davis MM, Beck AF, Simmons JM. Medical home quality and readmission risk for children hospitalized with asthma exacerbations. Pediatrics 2013; 131:64-70.
3. Beck AF, Simmons JM, Huang B, Kahn RS. Geomedicine: Area-based socioeconomic measures for assessing risk of hospital reutilization among children admitted for asthma. Am J Public Health 2012; 102:2308-2314.
4. Berry JG. 30-Day Pediatric Readmissions: Measurement and Prevention. Available at: www.ehcca.com/presentations/readsummit3/berry_ms4.pdf. Accessed October 27, 2014.
5. Berry JG, Toomey SL, Zaslavsky AM, Jha AK, Nakamura MM, Klein DJ, Feng JY, Shulman S, Chiang VW, Kaplan W, Hall M, Schuster MA. Pediatric readmission prevalence and variability across hospitals. JAMA 2013; 309:372-380. PMC3640861.

6. Bloomberg GR, Trinkaus KM, Fisher EB, Jr., Musick JR, Strunk RC. Hospital readmissions for childhood asthma: A 10-year metropolitan study. Am J Respir Crit Care Med 2003; 167:1068-1076.
7. Brown JR, Chang CH, Zhou W, MacKenzie TA, Malenka DJ, Goodman DC. Health system characteristics and rates of readmission after acute myocardial infarction in the United States. Journal of the American Heart Association. 2014 Jun 27;3(3):e000714.
8. Center for Medicare & Medicaid Services. Readmissions Reduction Program. 2013. Available at: http://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/AcuteInpatientPPS/Readmissions-Reduction-Program.html. Accessed January 2, 2014.
9. Coller RJ, Klitzner TS, Lerner CF, Chung PJ. Predictors of 30-day readmission and association with primary care follow-up plans. J Pediatr 2013; 163:1027-1033.
10. Coller RJ, Nelson BB, Sklansky DJ, Saenz AA, Klitzner TS, Lerner CF, Chung PJ. Preventing hospitalizations in children with medical complexity: a systematic review. Pediatrics. 2014 Nov 1:peds-2014.
11. Czaja AS, Hosokawa PW, Henderson WG. Unscheduled readmissions to the PICU: Epidemiology, risk factors, and variation among centers. Pediatr Crit Care Med 2013; 14:571- 579.
12. Escobar GJ, Joffe S, Gardner MN, Armstrong MA, Folck BF, Carpenter DM. Rehospitalization in the first two weeks after discharge from the neonatal intensive care unit. Pediatrics 1999; 104:e2.
13. Herrin J, St Andre J, Kenward K, Joshi MS, Audet AM, Hines SC. Community factors and hospital readmission rates. Health services research. 2015 Feb 1;50(1):20-39.
14. Hoehn RS, Wima K, Vestal MA, Weilage DJ, Hanseman DJ, Abbott DE, Shah SA. Effect of hospital safety-net burden on cost and outcomes after surgery. JAMA surgery. 2015 Oct 14:1-9.
15. Horwitz LI, Lin Z, Herrin J, Bernheim S, Drye EE, Krumholz HM, Ross JS. Association of hospital volume with readmission rates: a retrospective cross-sectional study. bmj. 2015 Feb 9;350:h447.
16. Joynt K, Jha AK. WHICH HOSPITALS HAVE HIGHER READMISSION RATES FOR PATIENTS WITH HEART FAILURE?. Journal of the American College of Cardiology. 2010 Mar 9;55(10):A143-E1348
17. Joynt KE, Jha AK. Who has higher readmission rates for heart failure, and why? Implications for efforts to improve care using financial incentives. Circulation: Cardiovascular Quality and Outcomes. 2011 Jan 1;4(1):53-9.
18. Krumholz, Harlan M., et al. "Relationship between hospital readmission and mortality rates for patients hospitalized with acute myocardial infarction, heart failure, or pneumonia." JAMA 309.6 2013; 587-593.
19. Li Z, Amstrong EJ, Parker JP, Danielsen B, Romano PS. Hospital variation in readmission after coronary artery bypass surgery in California. Circulation: Cardiovascular Quality and Outcomes. 2012 Sep 1;5(5):729-37.
20. Liu SY, Pearlman DN. Hospital readmissions for childhood asthma: The role of individual and neighborhood factors. Public Health Rep 2009; 124:65-78. PMC2602932.
21. Lorch SA, Baiocchi M, Silber JH, Even-Shoshan O, Escobar GJ, Small DS. The role of outpatient facilities in explaining variations in risk-adjusted readmission rates between hospitals. Hlth Serv Res 2010; 45:24-41. PMC2813435.
22. Lorch SA. Quality measurements in pediatrics: what do they assess?. JAMA pediatrics. 2013 Jan 1;167(1):89-90.
23. Lorch SA, Passarella M, Zeigler A. Challenges to measuring variation in readmission rates of neonatal intensive care patients. Acad Pediatr 2014; 14:S47-S53.
24. Martens PJ, Derksen S, Gupta S. Predictors of hospital readmission of Manitoba newborns within six weeks postbirth discharge: A population-based study. Pediatrics 2004; 114:708-713.
25. McGregor MJ, Reid RJ, Schulzer M, Fitzgerald JM, Levy AR, Cox MB. Socioeconomic status and hospital utilization among younger adult pneumonia admissions at a Canadian hospital. BMC Health Serv Res 2006; 6:152. PMC1697815.
26. McMillan JE, Meier ER, Winer JC, Coco M, Daymont M, Long S, Jacobs BR. Clinical and Geographic Characterization of 30-Day Readmissions in Pediatric Sickle Cell Crisis Patients. Hospital pediatrics. 2015 Aug 1;5(8):423-31.
27. Misky GJ, Wald HL, Coleman EA. Post-hospitalization transitions: Examining the effects of timing of primary care provider follow-up. Journal of Hospital Medicine. 2010 Sep 1;5(7):392-7.
28. Morse RB, Hall M, Fieldston ES, McGwire G, Anspacher M, Sills MR, Williams K, Oyemwense N, Mann KJ, Simon HK, Shah SS. Hospital-level compliance with asthma care quality measures at children's hospitals and subsequent asthma-related outcomes. JAMA 2011; 306:1454-1460.
29. Morris BH, Gard CC, Kennedy K. Rehospitalization of extremely low birth weight (ELBW) infants: Are there racial/ethnic disparities? J Perinatol 2005; 25:656-663.

30. Nakamura, Mari M., et al. "Measuring pediatric hospital readmission rates to drive quality improvement." Academic pediatrics 14.5 2014; S39-S46.
31. Phibbs CS, Baker LC, Caughey AB, Danielsen B, Schmitt SK, Phibbs RH. Level and volume of neonatal intensive care and mortality in very-low-birth-weight infants. New England Journal of Medicine. 2007 May 24;356(21):2165-75.
32. Phibbs CS, Bronstein JM, Buxton E, Phibbs RH. The effects of patient volume and level of care at the hospital of birth on neonatal mortality. Jama. 1996 Oct 2;276(13):1054-9.
33. Ray KN, A. LS. Hospitalization of early preterm, late preterm, and term infants during the first year of life by gestational age. Hospital Pediatrics 2013; 3:194-203.
34. Ray KN, Lorch SA. Hospitalization of rural and urban infants during the first year of life. Pediatrics 2012; 130:1084-1093.
35. Rice-Townsend S, Hall M, Barnes JN, Lipsitz S, Rangel SJ. Variation in risk-adjusted hospital readmission after treatment of appendicitis at 38 children's hospitals: An opportunity for collaborative quality improvement. Ann Surg 2013; 257:758-765.
36. Srivastava R, Keren R. Pediatric readmissions as a hospital quality measure. JAMA 2013; 309:396-398.
37. Tsai TC, Joynt KE, Orav EJ, Gawande AA, Jha AK. Variation in surgical-readmission rates and quality of hospital care. N Engl J Med 2013; 369:1134-1142.
38. Underwood MA, Danielsen B, Gilbert WM. Cost, causes and rates of rehospitalization of preterm infants. J Perinatol 2007; 27:614-619.
39. Wade KC, Lorch SA, Bakewell-Sachs S, Medoff-Cooper B, Silber JH, Escobar GJ. Pediatric care for preterm infants after NICU discharge: High number of office visits and prescription medications. J Perinatol 2008; 28:696-701.
40. Yeh RW, Rosenfield K, Zelevinsky K, Mauri L, Sakhuja R, Shivapour DM, Lovett A, Weiner BH, Jacobs AK, Normand SL. Sources of hospital variation in short-term readmission rates after percutaneous coronary intervention. Circulation: Cardiovascular Interventions. 2012 Apr 1;5(2):227-36.
41. Yu H, Wier LM, Elixhauser A. Hospital Stays for Children, 2009. HCUP Statistical Brief #118. Rockville, MD: Agency for Healthcare Research and Quality; August 2011. Available at: http://www.hcup-us.ahrq.gov/reports/statbriefs/sb118.jsp. Accessed January 29, 2016.

Note: For health outcome/PRO performance measures, no further information is required; however, you may provide evidence for any of the structures, processes, interventions, or service identified above.

INTERMEDIATE OUTCOME, PROCESS, OR STRUCTURE PERFORMANCE MEASURE

1a.3. Briefly state or diagram the path between structure, process, intermediate outcome, and health outcomes. Include all the steps between the measure focus and the health outcome.

1a.3.1. What is the source of the systematic review of the body of evidence that supports the performance measure?

☐ Clinical Practice Guideline recommendation – ***complete sections 1a.4, and 1a.7***

☐ US Preventive Services Task Force Recommendation – ***complete sections 1a.5 and 1a.7***

☐ Other systematic review and grading of the body of evidence (*e.g., Cochrane Collaboration, AHRQ Evidence Practice Center*) – ***complete sections 1a.6 and 1a.7***

☐ Other – ***complete section 1a.8***

Please complete the sections indicated above for the source of evidence. You may skip the sections that do not apply.

1a.4. CLINICAL PRACTICE GUIDELINE RECOMMENDATION

1a.4.1. Guideline citation (*including date*) and **URL for guideline** (*if available online*):

1a.4.2. Identify guideline recommendation number and/or page number and **quote verbatim, the specific guideline recommendation**.

1a.4.3. Grade assigned to the quoted recommendation with definition of the grade:

1a.4.4. Provide all other grades and associated definitions for recommendations in the grading system. (*Note: If separate grades for the strength of the evidence, report them in section 1a.7.*)

1a.4.5. Citation and URL for methodology for grading recommendations (*if different from 1a.4.1*)**:**

1a.4.6. If guideline is evidence-based (rather than expert opinion), are the details of the quantity, quality, and consistency of the body of evidence available (e.g., evidence tables)?

☐ Yes → ***complete section 1a.7***

☐ No → ***report on another systematic review of the evidence in sections 1a.6 and 1a.7; if another review does not exist, provide what is known from the guideline review of evidence in 1a.7***

1a.5. UNITED STATES PREVENTIVE SERVICES TASK FORCE RECOMMENDATION

1a.5.1. Recommendation citation (*including date*) and **URL for recommendation** (*if available online*):

1a.5.2. Identify recommendation number and/or page number and **quote verbatim, the specific recommendation**.

1a.5.3. Grade assigned to the quoted recommendation with definition of the grade:

1a.5.4. Provide all other grades and associated definitions for recommendations in the grading system. (*Note: the grading system for the evidence should be reported in section 1a.7.*)

1a.5.5. Citation and URL for methodology for grading recommendations (*if different from 1a.5.1*)**:**

Complete section 1a.7

1a.6. OTHER SYSTEMATIC REVIEW OF THE BODY OF EVIDENCE

1a.6.1. Citation (*including date*) and **URL** (*if available online*):

1a.6.2. Citation and URL for methodology for evidence review and grading (*if different from 1a.6.1*)**:**

Complete section 1a.7

1a.7. FINDINGS FROM SYSTEMATIC REVIEW OF BODY OF THE EVIDENCE SUPPORTING THE MEASURE

If more than one systematic review of the evidence is identified above, you may choose to summarize the one (or more) for which the best information is available to provide a summary of the quantity, quality, and consistency of the body of evidence. Be sure to identify which review is the basis of the responses in this section and if more than one, provide a separate response for each review.

1a.7.1. What was the specific structure, treatment, intervention, service, or intermediate outcome addressed in the evidence review?

1a.7.2. Grade assigned for the quality of the quoted evidence with definition of the grade:

1a.7.3. Provide all other grades and associated definitions for strength of the evidence in the grading system.

1a.7.4. What is the time period covered by the body of evidence? (*provide the date range, e.g., 1990-2010*). Date range: Click here to enter date range

QUANTITY AND QUALITY OF BODY OF EVIDENCE

1a.7.5. How many and what type of study designs are included in the body of evidence? (*e.g., 3 randomized controlled trials and 1 observational study*)

1a.7.6. What is the overall quality of evidence <u>across studies</u> in the body of evidence? (*discuss the certainty or confidence in the estimates of effect particularly in relation to study factors such as design flaws, imprecision due to small numbers, indirectness of studies to the measure focus or target population*)

ESTIMATES OF BENEFIT AND CONSISTENCY ACROSS STUDIES IN BODY OF EVIDENCE

1a.7.7. What are the estimates of benefit—magnitude and direction of effect on outcome(s) <u>across studies</u> in the body of evidence? (*e.g., ranges of percentages or odds ratios for improvement/ decline across studies, results of meta-analysis, and statistical significance*)

1a.7.8. What harms were studied and how do they affect the net benefit (benefits over harms)?

UPDATE TO THE SYSTEMATIC REVIEW(S) OF THE BODY OF EVIDENCE

1a.7.9. If new studies have been conducted since the systematic review of the body of evidence, provide for <u>each</u> new study: 1) citation, 2) description, 3) results, 4) impact on conclusions of systematic review.

1a.8 OTHER SOURCE OF EVIDENCE

If source of evidence is NOT from a clinical practice guideline, USPSTF, or systematic review, please describe the evidence on which you are basing the performance measure.

1a.8.1 What process was used to identify the evidence?

1a.8.2. Provide the citation and summary for each piece of evidence.

1. Evidence, Performance Gap, Priority – Importance to Measure and Report

Extent to which the specific measure focus is evidence-based, important to making significant gains in healthcare quality, and improving health outcomes for a specific high-priority (high-impact) aspect of healthcare where there is variation in or overall less-than-optimal performance. ***Measures must be judged to meet all subcriteria to pass this criterion and be evaluated against the remaining criteria.***

1a. Evidence to Support the Measure Focus – See attached Evidence Submission Form
NQF_Evidence_NICU_Readmissions_30_v1.docx

1b. Performance Gap

Demonstration of quality problems and opportunity for improvement, i.e., data demonstrating:

- considerable variation, or overall less-than-optimal performance, in the quality of care across providers; and/or
- disparities in care across population groups.

1b.1. **Briefly explain the rationale for this measure** *(e.g., the benefits or improvements in quality envisioned by use of this measure)*
Hospital readmissions have been an area of particular interest for state and national policy agencies, health insurers, and caregivers because of the high costs, both financial and to families, associated with them. So-called "preventable" readmissions, described because ostensibly some change in practice at either the inpatient or outpatient level could have prevented the readmission, may provide insight into care practices that could limit hospitalizations. In pediatric medicine, there are groups of high-risk patients for whom hospital readmissions occur frequently. For example, while the estimated readmission rate within 30 days among the 2.4 million admissions annually in the United States is approximately 6.5% (Berry 2013, Yu 2011), many conditions such as surgery, sickle cell disease, and prematurity have rates between 15 and 20% (Ray 2013, Wade 2008, Underwood 2007, Berry 2014). One other group is children discharged from the neonatal intensive care unit (NICU). Premature infants have an approximately three-fold increase in risk of hospital readmission after discharge compared to term infants, with higher rates in infants of younger gestational age (Ray 2013). These hospital readmissions contribute to the higher health care costs and utilization seen in prematurely-born infants (Wade 2008, Underwood 2007). For the preterm population, we have demonstrated substantial variation in the unadjusted readmission rates among California hospitals: Mean rate of 5.8%; median of 4.5%; Interquartile range 2.5%-7.8%; Range 0%-50%; Standardized difference 909% for the range, and 97% for the interquartile range. The large variation between hospitals persisted after adjusting for gestational age and sociodemographic factors: mean rate 6.0%; median 4.8%; interquartile range 2.5%-7.9%; range 0%-52.8%; standardized difference 943% for the range, and 96% for the interquartile range. Of note, the average percent change between the unadjusted and risk-adjusted rates was 0.75% (that is, the risk-adjusted rate was less than 1% different from the unadjusted rate; for the median unadjusted rate of 4.5%, this would result in a risk-adjusted rate of 4.545%), with a range of -3.4% to 5.6%.
Similar differences are seen between states within their Medicaid population, suggesting that larger systems of care beyond the NICU itself may be important to optimize the health of infants who require neonatal intensive care. These systems may be at the health system level or at the level of the outpatient care provider (Lorch 2010).
Based on several conceptual frameworks for readmissions in the pediatric or specific neonatal population, readmission rates can be considered a product of:
- patient factors, such as illness severity at the time of discharge and socioeconomic factors at the patient and community level, such as access to care and health insurance; and
- inpatient quality of care provided during the hospitalization, which may impact readmissions either directly, or more likely through altering the severity of illness of a specific patient at the time of discharge through differing rates of bronchopulmonary dysplasia, necrotizing entercolitis, or intraventricular hemorrhage, each of which has been associated with higher readmission rates in prematurely-born infants (Lorch 2010, Ray 2012, Lorch 2014); and
- care processes around the time of discharge, including education of families to prepare them for discharge and processes to ensure the transition of care from the inpatient to the outpatient setting; and
- outpatient care quality (Lorch 2010).
Therefore, implementing a quality measure around readmission rates from the neonatal intensive care unit will result in improved care around the transition from inpatient to outpatient care; strengthen the post-discharge support for the infants who required neonatal intensive care; and decrease overall costs of health care association with variation in rehospitalization rates of these infants, especially if the readmission measure assesses a different aspect of quality from other potential measures, such as complications of preterm birth. This assessment will be demonstrated with poor correlation between readmission rates and hospital complication rates, both risk adjusted and unadjusted.

1b.2. **Provide performance scores on the measure as specified (current and over time) at the specified level of analysis.** *(This is*

required for endorsement maintenance. Include mean, std dev, min, max, interquartile range, scores by decile. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included). This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.

We performed a risk adjustment on a cohort of infants born in California at a gestational age between 23 and 34 weeks between 1995 and 2009. The department of health linked these infants' birth certificates to death certificates using name and date of birth, and then de-identified the records. Then, over 98% of these records were linked to maternal and newborn hospital records using prior methods (Herrchen 1997, Phibbs 2007). Over 80% of the unmatched live birth or fetal death certificate records were missing the delivery hospital, suggesting a birth at home or a birthing center. The unmatched records had similar gestational age and racial/ethnic distributions to the matched records. Because this dataset contains records for all hospitals in California, we can measure readmissions at any California hospital, not simply readmissions to the discharging hospital. To ensure that there were enough patients per hospital to make reliable estimates of the readmission rate, (Silber 2010) we limited the analyses to those hospitals who discharged over 50 eligible patients per year (N=154).

We initially considered using an unadjusted measure, paralleling readmission projects from other groups such as Medicare that consider any readmission a potential quality of care issue regardless of the reason. However, based on data from our work and others, we included specific infant characteristics that are known to influence readmission risk. The inclusion of these data elements would improve face validity of the measure. Thus, for risk adjustment, we include characteristics of the infant that may increase the risk of hospital readmission after discharge from the NICU based on prior work: gestational age, birth weight, gender, and insurance status (Ray 2013, Lorch 2010, Lorch 2012). Gestational age and birth weight are specifically captured in birth certificate records. We also assessed how risk-adjusted hospital rates changed when we included common complications of preterm birth associated with readmissions in prior work, and captured using ICD-9CM codes in hospital administrative records: bronchopulmnary dysplasia (BPD), necrotizing enterocolitis (NEC), retinopathy of prematurity (ROP), and intraventricular hemorrhage (IVH). Including these factors is controversial when assessing a facility's quality, though, because inpatient quality of care may result in higher complication rates. Gestational age, birth weight, sociodemographic information, and complications of premature birth were available in over 98% of records in the California state data, and have been used in prior work from this dataset (Phibbs 2007, Lorch 2012).

The California data provides a large dataset with an extensive number of hospitals to examine these questions (Table 1 in the Appendix). There were an average of 67-83 patients/hospital annually with a wide variation in numbers. The number of level 2+ hospitals rose from 37.38% in 1995 to 47.62% in 2009, mostly through increases in level 3a and 3b hospitals.

Adjusted rates of readmissions over a 1 year time period varied widely among these hospitals (section 2b5 of the testing form and Table 3 in the Appendix). In fact, the addition of common complications of preterm birth to the risk adjustment model made little difference to the readmission rates, with an average relative increase of 0.08% (standard deviation 3.31%) for readmissions 30 days after discharge.

1b.3. If no or limited performance data on the measure as specified is reported in 1b2, then provide a summary of data from the literature that indicates opportunity for improvement or overall less than optimal performance on the specific focus of measurement.

N/A

1b.4. Provide disparities data from the measure as specified (current and over time) by population group, e.g., by race/ethnicity, gender, age, insurance status, socioeconomic status, and/or disability. *(This is required for endorsement maintenance. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities include.) This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*

The CA data was utilized to test for disparities in NICU readmission rates at the state-level. For these analyses, data for the entire state population are presented for those patients discharged from a hospital included in the readmission rate analysis.

Race/ethnicity (Appendix Table 4): For these analyses, race and ethnicity was determined based on the race/ethnicity variable reported in the data. White was defined as white, not of Hispanic origin. Black was defined as black, not of Hispanic origin. For Hispanic, we combined children reported as ´Hispanic or Latino´ and ´Hispanic or Latino and one or more races´. Other included American Indian, Alaskan Native, Asian, Pacific Islander and children with missing race/ethnicity. We stratified the Readmissions metric by enrollee race/ethnicity. Readmission rates did not vary substantially between race/ethnicities at the state level. This is similar to previous work from our group finding no real difference in readmission rates between children of different racial/ethnic backgrounds (Ray 2003).

Special Health Care Needs (Appendix Table 5): Based on published peer-reviewed literature, we compiled a list of pediatric chronic conditions (ensuing list of ICD-9 codes) where each condition was represented in all or most of the papers (Valentine, 2000; Ireys, 1997; Todd, 2006; Fowler, 2001; Neuzil, 2000; Feudtner, 2000; Feudtner, 2001; Seferian, 2006). Readmissions by Special Health Care Needs were tested. Unsurprisingly, children with special healthcare needs were more likely than health children to have a readmission.

Socioeconomic Status (Appendix Table 6): Socioeconomic measures at the individual or census-tract level are not included in the CA data. Although 5-digit-zip code-based socioeconomic measures have significant limitations, we performed analyses using two socioeconomic variables (% with high school degree and income level) stratified by quartiles in order to demonstrate that these analyses are feasible (Krieger, 1997). These variables were abstracted from U.S. census 5-digit-zip code-level data and merged with the data. If 9-digit-zip code data were available in the CA data, these analyses would produce more robust and meaningful results. As noted in the methods, these analyses were performed for the purposes of demonstrating feasibility and NOT for the purposes of assessing the significance of associations. Although there was an association between higher socioeconomic status and lower readmissions, the difference was low.

Rurality/Urbanicity (Appendix Table 8): A crosswalk was performed between the CA data using the 2010 Census urban and rural classification (http://www.census.gov/geo/www/ua/2010urbanruralclass.html). There are two types of urban areas: urbanized areas have 50,000 or more people residing in that area; urban clusters have at least 2,500 and less than 50,000 people residing in that area. Rural area encompasses all population, housing, and territory not included within an urban area. In general, there was relatively little variation between the geographic categories overall.

1b.5. **If no or limited data on disparities from the measure as specified is reported in 1b4, then provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement. Include citations.**
Not necessary if performance data provided in 1b.4

1c. High Priority (previously referred to as High Impact)

The measure addresses:

- a specific national health goal/priority identified by DHHS or the National Priorities Partnership convened by NQF; OR
- a demonstrated high-priority (high-impact) aspect of healthcare (e.g., affects large numbers of patients and/or has a substantial impact for a smaller population; leading cause of morbidity/mortality; high resource use (current and/or future); severity of illness; and severity of patient/societal consequences of poor quality).

1c.1. **Demonstrated high priority aspect of healthcare**
Affects large numbers, High resource use, Severity of illness
1c.2. **If Other:**

1c.3. **Provide epidemiologic or resource use data that demonstrates the measure addresses a high priority aspect of healthcare. List citations in 1c.4.**
Preterm births account for 12% of all live births in the United States, with rates in non-Hispanic black patients being 50-60% higher than other racial/ethnic groups. Readmissions are particularly prevalent among premature infants, who have been shown to have an approximately three-fold increase in risk of hospital admissions after discharge across all time frames compared to term infants, with higher rates in infants of younger gestational age (Ray 2013, Escobar 1999). Infants with complications or other health conditions such as bronchopulmonary dysplasia and necrotizing enterocolitis also experience higher rates of readmission (Morris 2005). Issues surrounding NICU readmissions are particularly relevant to the African American community, as a larger proportion of babies born to black mothers are premature, even after adjusting for income, education level, and socioeconomic status (Morris 2005). Additional studies have found higher rates of readmission among non-Hispanic black and Hispanic babies (Kuzniewicz 2013, Berry 2013). Additionally, the increased prevalence of jaundice in the Asian population likewise increases their risk of readmission (Kuzniewicz 2013, Paul 2006). Preventing hospital readmissions is an area of emphasis by insurers and public health professionals, because hospital readmissions may represent either poor quality of care during the hospitalization, or poor discharge planning and transition of care from inpatient to outpatient providers (Tsai 2013, Berry 2013, Morse 2011, Escobar 2005, Profit 2007, Lorch 2010). Readmissions can be used to define or measure the effectiveness of infant discharge criteria (Seki 2011, Kotagal 1995), or the effect of performance-based quality metrics (Paul 2006). Analysis of readmissions on longer time intervals can also be used to assess quality of outpatient care (Lorch 2010). Increasing access to and continuity of health insurance for mothers and infants, as well as identification of maternal risk factors (young age, first-time pregnancy, diabetes, hypertension, etc.) and targeting mothers for special education can also potentially reduce readmission rates (Paul 2006, Bakewell-Sachs 2004, Ray 2010).
The costs and stresses of an infant admitted to the NICU can have a profound effect on family well-being. Several studies have found elevated levels of hostility, anxiety, and/or depression among parents of NICU infants. (Doering 2000, Carter 2005) These alterations in parental attitudes and family well-being can produce long-term effects on the development of the child and family. Caring for a premature infant also requires more maternal/family education, failure of which can further increase risk of readmission (Bakewell-Sachs 2004, Paul 2006). Increased hospitalizations contribute to higher healthcare costs and utilization (Wade 2008, Kirkby 2007). Costs and resource utilization by preterm, low birth weight infants (those at the highest risk of readmission) are substantially higher

(according to Gilbert, et. al. $224,000 at 500-600 g, vs. $1000 at 3000g or greater) (Gilbert 2003, Russell 2007). Although the initial NICU admission is the highest-cost, each readmission has an average claim of approximately $8,468 using data from the early 2000's (Underwood 2007). Premature infants and infants with morbidities have been shown to have a higher number of office visits (especially for higher-cost non-well child visits) and a greater number of prescriptions (Wade 2008).

Finally, increased risk of social and behavioral problems associated with prematurity can have lingering effects over the entire life of the child. Early pediatric interventions have been shown to reduce these risks. Although readmissions are not themselves associated with a child's future health, they are more common among infants with health problems that require special healthcare attention, such as prematurity, low birth weight, and other neonatal morbidities. Readmissions as a measure can help ensure that these babies are receiving the routine and preventive care necessary to improve their health outcomes, as quality outpatient and primary pediatric care will reduce preventable readmissions. Premature infants and infants with morbidities have been shown to have delayed achievement of physiologic milestones such as respiration and feeding (Bakewell-Sachs 2009). In multiple studies, including multi-study reviews, of outcomes for babies born preterm versus term, preterm infants had significantly lower cognitive scores, educational ability, and need for medical interventions, as well as an increased relative risk of developing ADHD (Bhutta 2002, Chapieski 1997, McGowan 2010). Several programs aimed at early intervention aimed at reducing the developmental delay of preterm infants via parental education, family support, and pediatric follow-up have shown improved cognitive scores (Brooks-Gunn 1993).

Overall, then, the high rate of readmissions in prematurely-born infants and infants who require neonatal intensive care at birth, along with the substantial interhospital and interstate variation in these rates, mirror the concerns faced by caregivers of adults and older children when readmission rates of these populations became a high priority area of the health care system.

1c.4. Citations for data demonstrating high priority provided in 1a.3

Bakewell-Sachs, Susan, and Susan Gennaro. "Parenting the post-NICU premature infant." MCN: The American Journal of Maternal/Child Nursing 29.6 (2004): 398-403.

Bakewell-Sachs, Susan, et al. "Infant functional status: the timing of physiologic maturation of premature infants." Pediatrics 123.5 (2009): e878-e886.

Berry JG, Toomey SL, Zaslavsky AM, et al. Pediatric readmission prevalence and variability across hospitals. JAMA. 2013;309:372-380.

Bhutta AT, Cleves MA, Casey PH, Cradock MM, Anand KS. Cognitive and Behavioral Outcomes of School-Aged Children Who Were Born Preterm: A Meta-analysis. JAMA.2002;288(6):728-737.

Brooks-Gunn, J., Klebanov, P. K., Liaw, F.-r. and Spiker, D. (1993), Enhancing the Development of Low-Birthweight, Premature Infants: Changes in Cognition and Behavior over the First Three Years. Child Development, 64: 736–753.

Carter, J. D., et al. "Infants in a neonatal intensive care unit: parental response."Archives of Disease in Childhood-Fetal and Neonatal Edition 90.2 (2005): F109-F113.

Chapieski, M. Lynn, and Karen D. Evankovich. "Behavioral effects of prematurity." Seminars in perinatology. Vol. 21. No. 3. WB Saunders, 1997.

Doering, Lynn V., Debra K. Moser, and Kathleen Dracup. "Correlates of anxiety, hostility, depression, and psychosocial adjustment in parents of NICU infants."Neonatal Network: The Journal of Neonatal Nursing 19.5 (2000): 15-23.

Escobar GJ, Greene JD, Hulac P, et al. Rehospitalisation after birth hospitalisation: Patterns among infants of all gestations. Arch Dis Child. 2005;90:125-131.

Escobar, Gabriel J., et al. "Rehospitalization in the first two weeks after discharge from the neonatal intensive care unit." Pediatrics 104.1 (1999): e2-e2.

Gilbert, William M., Thomas S. Nesbitt, and Beate Danielsen. "The cost of prematurity: quantification by gestational age and birth weight." Obstetrics & Gynecology 102.3 (2003): 488-492.

Kirkby, Sharon, et al. "Clinical outcomes and cost of the moderately preterm infant." Advances in Neonatal Care 7.2 (2007): 80-87.

Kotagal, Uma R., et al. "Description and evaluation of a program for the early discharge of infants from a neonatal intensive care unit." The Journal of pediatrics 127.2 (1995): 285-290.

Kuzniewicz, Michael W., et al. "Hospital readmissions and emergency department visits in moderate preterm, late preterm, and early term infants."Clinics in perinatology 40.4 (2013): 753-775.

Lorch SA, Baiocchi M, Silber JH, et al. The role of outpatient facilities in explaining variations in risk-adjusted readmission rates between hospitals. Hlth Serv Res. 2010;45:24-41.

McGowan, Jennifer E., et al. "Early childhood development of late-preterm infants: a systematic review." Pediatrics (2011): peds-2010.

Morris BH, Gard CC, Kennedy K. Rehospitalization of extremely low birth weight (ELBW) infants: Are there racial/ethnic disparities? J Perinatol. 2005;25:656-663.

Morse RB, Hall M, Fieldston ES, et al. Hospital-level compliance with asthma care quality measures at children´s hospitals and subsequent asthma-related outcomes. JAMA. 2011;306:1454-1460.

Paul, Ian M., et al. "Preventable newborn readmissions since passage of the Newborns' and Mothers' Health Protection Act." Pediatrics 118.6 (2006): 2349-2358.

Profit J, McCormick MC, Escobar GJ, et al. Neonatal intensive care unit census influences discharge of moderately preterm infants. Pediatrics. 2007;119:314-319.

Ray K, Escobar GJ, Lorch SA. Premature infants born to adolescent mothers: Health care utilization after initial discharge academic pediatrics. Acad Pediatr. 2010;10:302-308.

Ray KN, A. LS. Hospitalization of early preterm, late preterm, and term infants during the first year of life by gestational age. Hospital Pediatrics. 2013;3:194-203.

Russell, Rebecca B., et al. "Cost of hospitalization for preterm and low birth weight infants in the United States." Pediatrics 120.1 (2007): e1-e9.

Seki, Kazuo, et al. "Early discharge from a neonatal intensive care unit and rates of readmission." Pediatrics International 53.1 (2011): 7-12.

Tsai TC, Joynt KE, Orav EJ, et al. Variation in surgical-readmission rates and quality of hospital care. N Engl J Med. 2013;369:1134-1142.

Underwood MA, Danielsen B, Gilbert WM. Cost, causes and rates of rehospitalization of preterm infants. J Perinatol. 2007;27:614-619.

Wade KC, Lorch SA, Bakewell-Sachs S, et al. Pediatric care for preterm infants after NICU discharge: High number of office visits and prescription medications. J Perinatol. 2008;28:696-701.

1c.5. If a PRO-PM (e.g. HRQoL/functional status, symptom/burden, experience with care, health-related behaviors), provide evidence that the target population values the measured PRO and finds it meaningful. *(Describe how and from whom their input was obtained.)*
N/A

2. Reliability and Validity—Scientific Acceptability of Measure Properties

Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. ***Measures must be judged to meet the subcriteria for both reliability and validity to pass this criterion and be evaluated against the remaining criteria.***

2a.1. Specifications The measure is well defined and precisely specified so it can be implemented consistently within and across organizations and allows for comparability. eMeasures should be specified in the Health Quality Measures Format (HQMF) and the Quality Data Model (QDM).

De.5. **Subject/Topic Area** *(check all the areas that apply):*
Perinatal and Reproductive Health : Newborn

De.6. **Cross Cutting Areas** *(check all the areas that apply):*
Care Coordination : Readmissions, Safety : Readmissions

S.1. **Measure-specific Web Page** *(Provide a URL link to a web page specific for this measure that contains current detailed specifications including code lists, risk model details, and supplemental materials. Do not enter a URL linking to a home page or to general information.)*
N/A

S.2a. **If this is an eMeasure**, HQMF specifications must be attached. Attach the zipped output from the eMeasure authoring tool (MAT) - if the MAT was not used, contact staff. (Use the specification fields in this online form for the plain-language description of the specifications)
This is not an eMeasure **Attachment:**

S.2b. **Data Dictionary, Code Table, or Value Sets** *(and risk model codes and coefficients when applicable) must be attached. (Excel or csv file in the suggested format preferred - if not, contact staff)*
Attachment **Attachment:** Data_Dictionary-635948697097724496.xlsx

S.3. **For endorsement maintenance**, please briefly describe any changes to the measure specifications since last endorsement date and explain the reasons.
N/A

S.4. **Numerator Statement** *(Brief, narrative description of the measure focus or what is being measured about the target population, i.e., cases from the target population with the target process, condition, event, or outcome)*
IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
Number of infants with a gestational age between 23-34 weeks who were readmitted to the hospital within 30 days of discharge. These time periods are assessed cumulatively, such that readmissions occurring within prior time periods are included. Reliability is strongest if each health care unit has at least 50 discharges per time unit studied.

S.5. **Time Period for Data** *(What is the time period in which data will be aggregated for the measure, e.g., 12 mo, 3 years, look back to August for flu vaccination? Note if there are different time periods for the numerator and denominator.)*
Up to 30 days post discharge. This metric may assess care in the outpatient setting (Lorch 2010), but the test characteristics of the metric, particularly inter-year reliability and model calibration, lead to concerns about its ability to distinguish care delivered during the neonatal intensive care unit time period. Further data will be needed to validate the measure this far from neonatal intensive care unit discharge.

S.6. **Numerator Details** *(All information required to identify and calculate the cases from the target population with the target process, condition, event, or outcome such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
IF an OUTCOME MEASURE, describe how the observed outcome is identified/counted. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
Number of eligible newborns with an inpatient readmission within 30 days of discharge, who survive to time of hospital discharge. The optimal measure is risk-adjusted using gestational age, race, gender, education, insurance status, and complications (bronchopulmnary dysplasia (BPD), necrotizing enterocolitis (NEC), retinopathy of prematurity (ROP), and intraventricular hemorrhage (IVH)).

S.7. **Denominator Statement** *(Brief, narrative description of the target population being measured)*
Number of newborns with a gestational age between 23-34 weeks discharged from the NICU, based on gestational age field contained in the birth certificate record (best obstetrical estimate).

S.8. **Target Population Category** *(Check all the populations for which the measure is specified and tested if any):*
Children's Health

S.9. **Denominator Details** *(All information required to identify and calculate the target population/denominator such as definitions, specific data collection items/responses , code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
Gestational age between 23 and 34 weeks as defined by the gestational age field in the vital statistics data.

S.10. **Denominator Exclusions** *(Brief narrative description of exclusions from the target population)*
Infants with a specified congenital anomaly are excluded from the target population.
Infants with a missing gestational age are excluded from the primary analysis. Information about multiple imputation methods to allow for their inclusion are presented in the testing attachment, section 2b7.
Infants who expired during the neonatal intensive care period are not eligible for a hospital readmission and excluded.
The smallest level of measurement (i.e. hospital, state, etc.) must have a minimum of 50 patients eligible for readmission in a single calendar year.

S.11. **Denominator Exclusion Details** *(All information required to identify and calculate exclusions from the denominator such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
Infants with a specified congenital anomaly are identified using the excel file in S.2b. Infants with a missing gestational age are excluded from the primary analysis, based on the gestational age field (best obstetrical estimate) from the birth certificate. Information about multiple imputation methods to allow for their inclusion are presented in the testing attachment, section 2b7.
Infants who expired will be identified using the outcome of the neonatal intensive care hospitalization.
The smallest level of measurement (i.e. hospital, state, etc.) must have a minimum of 50 patients eligible for readmission in a single calendar year.

S.12. **Stratification Details/Variables** *(All information required to stratify the measure results including the stratification variables, definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format with at S.2b)*
N/A

S.13. **Risk Adjustment Type** (Select type. Provide specifications for risk stratification in S.12 and for statistical model in S.14-15)
Statistical risk model
If other:

S.14. **Identify the statistical risk model method and variables** *(Name the statistical method - e.g., logistic regression and list all the risk factor variables. Note - risk model development and testing should be addressed with measure testing under Scientific Acceptability)*
The optimal model uses logistic regression is used to adjust for the risk factor variables, which are: adjusted for gestational age, race, gender, education, insurance status, and complications (bronchopulmonary dysplasia (BPD), necrotizing enterocolitis (NEC), retinopathy of prematurity (ROP), and intraventricular hemorrhage (IVH)).
If specific variables are not available for the risk adjustment model, gestational age or birth weight alone is the minimum set of variables needed for risk adjustment. This streamlined risk adjustment model will produce a reliable measure, albeit one with less face validity than models that include all variables listed above. Data suggest that the risk adjusted readmission rates change less than 0.3% when these factors are included in the model (Lorch 2014).

S.15. **Detailed risk model specifications** *(must be in attached data dictionary/code list Excel or csv file. Also indicate if available at measure-specific URL identified in S.1.)*
Note: Risk model details (including coefficients, equations, codes with descriptors, definitions), should be provided on a separate worksheet in the suggested format in the Excel or csv file with data dictionary/code lists at S.2b.

Provided in response box S.15a

S.15a. **Detailed risk model specifications** (*if not provided in excel or csv file at S.2b*)
The definitions are provided in the Excel file in S.2b, the risk model specifications are provided in the Testing Form, Table 16.

S.16. **Type of score:**
Rate/proportion
If other:

S.17. **Interpretation of Score** *(Classifies interpretation of score according to whether better quality is associated with a higher score, a lower score, a score falling within a defined interval, or a passing score)*
Better quality = Lower score

S.18. **Calculation Algorithm/Measure Logic** *(Describe the calculation of the measure score as an ordered sequence of steps including identifying the target population; exclusions; cases meeting the target process, condition, event, or outcome; aggregating data; risk adjustment; etc.)*
First, patients are identified as having a gestational age of 23-34 weeks using the best obstetrical estimate of gestational age included in the birth certificate data. Variables for adjustment are then determined: race, gender, education, insurance status, and complications (bronchopulmonary dysplasia (BPD), necrotizing enterocolitis (NEC), retinopathy of prematurity (ROP), and intraventricular hemorrhage (IVH)). For these patients, readmissions are identified within the specified time period of 30 days. We next calculate the observed number of readmissions divided by the expected number of readmissions at the given level of the health care provider (hospital, outpatient facility, state) where the expected number of readmissions is based on the risk adjustment model using the above variables. The final ratio, (observed number of readmissions)/(expected number of readmissions), is then multiplied by the mean percentage of admissions in the cohort to calculate a risk-adjusted readmission rate.
A similar method is performed if complications are excluded from the risk adjustment model. However, as we note, in testing of the various risk adjustment models the risk-adjusted readmission rates for each time period changed minimally when these complication measures were included or excluded from the model. The smallest level of measurement (i.e. hospital, state, etc.) must have a minimum of 50 patients eligible for readmission in a single calendar year.

S.19. **Calculation Algorithm/Measure Logic Diagram URL or Attachment** *(You also may provide a diagram of the Calculation Algorithm/Measure Logic described above at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
No diagram provided

S.20. **Sampling** *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*
IF a PRO-PM, identify whether (and how) proxy responses are allowed.
N/A

S.21. **Survey/Patient-reported data** *(If measure is based on a survey, provide instructions for conducting the survey and guidance on minimum response rate.)*
IF a PRO-PM, specify calculation of response rates to be reported with performance measure results.
N/A

S.22. **Missing data** (specify how missing data are handled, e.g., imputation, delete case.)
Required for Composites and PRO-PMs.
By the inclusion criteria, cases missing gestational age were not included in the analysis.

S.23. **Data Source** *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*
If other, please describe in S.24.
Administrative claims, Electronic Clinical Data : Electronic Health Record, Other

S.24. **Data Source or Collection Instrument** *(Identify the specific data source/data collection instrument e.g. name of database, clinical registry, collection instrument, etc.)*
IF a PRO-PM, identify the specific PROM(s); and standard methods, modes, and languages of administration.
N/A

S.25. **Data Source or Collection Instrument** *(available at measure-specific Web page URL identified in S.1 OR in attached appendix at*

A.1)
No data collection instrument provided

S.26. **Level of Analysis** *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*
Facility, Population : State

S.27. **Care Setting** *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*
Hospital/Acute Care Facility
If other:

S.28. **COMPOSITE Performance Measure** - Additional Specifications *(Use this section as needed for aggregation and weighting rules, or calculation of individual performance measures if not individually endorsed.)*
N/A

2a. Reliability – See attached Measure Testing Submission Form
2b. Validity – See attached Measure Testing Submission Form
NQF_Testing_NICU_Readmissions_30_v4_FINAL.docx

NATIONAL QUALITY FORUM—Measure Testing (subcriteria 2a2, 2b2-2b7)

Measure Number (*if previously endorsed*)**:** Click here to enter NQF number

Measure Title: Neonatal Intensive Care All-Condition Readmissions

Date of Submission: 3/29/2016

Type of Measure:

☐ Composite – ***STOP – use composite testing form***	☒ Outcome (*including PRO-PM*)
☐ Cost/resource	☐ Process
☐ Efficiency	☐ Structure

Instructions

- Measures must be tested for all the data sources and levels of analyses that are specified. ***If there is more than one set of data specifications or more than one level of analysis, contact NQF staff*** about how to present all the testing information in one form.
- **For all measures, sections 1, 2a2, 2b2, 2b3, and 2b5 must be completed.**
- **For outcome and resource use measures**, section **2b4** also must be completed.
- If specified for **multiple data sources/sets of specificaitons** (e.g., claims and EHRs), section **2b6** also must be completed.
- Respond to all questions as instructed with answers immediately following the question. All information on testing to demonstrate meeting the subcriteria for reliability (2a2) and validity (2b2-2b6) must be in this form. An appendix for *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 20 pages (*incuding questions/instructions;* minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.
- For information on the most updated guidance on how to address sociodemographic variables and testing in this form refer to the release notes for version 6.6 of the Measure Testing Attachment.

Note: The information provided in this form is intended to aid the Steering Committee and other stakeholders in understanding to what degree the testing results for this measure meet NQF's evaluation criteria for testing.

2a2. Reliability testing [10] demonstrates the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise. For **PRO-PMs and composite performance measures**, reliability should be demonstrated for the computed performance score.

2b2. Validity testing [11] demonstrates that the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality. For **PRO-PMs and composite performance measures**, validity should be demonstrated for the computed performance score.

2b3. Exclusions are supported by the clinical evidence; otherwise, they are supported by evidence of sufficient frequency of occurrence so that results are distorted without the exclusion; [12]
AND
If patient preference (e.g., informed decisionmaking) is a basis for exclusion, there must be evidence that the exclusion impacts performance on the measure; in such cases, the measure must be specified so that the information about patient

preference and the effect on the measure is transparent (e.g., numerator category computed separately, denominator exclusion category computed separately). [13]

2b4. For outcome measures and other measures when indicated (e.g., resource use):

- **an evidence-based risk-adjustment strategy** (e.g., risk models, risk stratification) is specified; is based on patient factors (including clinical and sociodemographic factors) that influence the measured outcome and are present at start of care; [14,15] and has demonstrated adequate discrimination and calibration

OR

- rationale/data support no risk adjustment/ stratification.

2b5. Data analysis of computed measure scores demonstrates that methods for scoring and analysis of the specified measure allow for **identification of statistically significant and practically/clinically meaningful [16] differences in performance**;

OR

there is evidence of overall less-than-optimal performance.

2b6. If multiple data sources/methods are specified, there is demonstration they produce comparable results.

2b7. For **eMeasures, composites, and PRO-PMs** (or other measures susceptible to missing data), analyses identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias.

Notes

10. Reliability testing applies to both the data elements and computed measure score. Examples of reliability testing for data elements include, but are not limited to: inter-rater/abstractor or intra-rater/abstractor studies; internal consistency for multi-item scales; test-retest for survey items. Reliability testing of the measure score addresses precision of measurement (e.g., signal-to-noise).

11. Validity testing applies to both the data elements and computed measure score. Validity testing of data elements typically analyzes agreement with another authoritative source of the same information. Examples of validity testing of the measure score include, but are not limited to: testing hypotheses that the measures scores indicate quality of care, e.g., measure scores are different for groups known to have differences in quality assessed by another valid quality measure or method; correlation of measure scores with another valid indicator of quality for the specific topic; or relationship to conceptually related measures (e.g., scores on process measures to scores on outcome measures). Face validity of the measure score as a quality indicator may be adequate if accomplished through a systematic and transparent process, by identified experts, and explicitly addresses whether performance scores resulting from the measure as specified can be used to distinguish good from poor quality.

12. Examples of evidence that an exclusion distorts measure results include, but are not limited to: frequency of occurrence, variability of exclusions across providers, and sensitivity analyses with and without the exclusion.

13. Patient preference is not a clinical exception to eligibility and can be influenced by provider interventions.

14. Risk factors that influence outcomes should not be specified as exclusions

15. With large enough sample sizes, small differences that are statistically significant may or may not be practically or clinically meaningful. The substantive question may be, for example, whether a statistically significant difference of one percentage point in the percentage of patients who received smoking cessation counseling (e.g., 74 percent v. 75 percent) is clinically meaningful; or whether a statistically significant difference of $25 in cost for an episode of care (e.g., $5,000 v. $5,025) is practically meaningful. Measures with overall less-than-optimal performance may not demonstrate much variability across providers.

1. DATA/SAMPLE USED FOR ALL TESTING OF THIS MEASURE

Often the same data are used for all aspects of measure testing. In an effort to eliminate duplication, the first five questions apply to all measure testing. If there are differences by aspect of testing,(e.g., reliability vs. validity) be sure to indicate the specific differences in question 1.7.

1.1. What type of data was used for testing? (*Check all the sources of data identified in the measure specifications and data used for testing the measure. Testing must be provided for all the sources of data specified and intended for measure implementation.* ***If different data sources are used for the numerator and denominator, indicate N [numerator] or D [denominator] after the checkbox.***)

Measure Specified to Use Data From: (*must be consistent with data sources entered in S.23*)	**Measure Tested with Data From:**
☐ abstracted from paper record	☐ abstracted from paper record
☒ administrative claims	☒ administrative claims
☐ clinical database/registry	☐ clinical database/registry
☒ abstracted from electronic health record	☒ abstracted from electronic health record
☐ eMeasure (HQMF) implemented in EHRs	☐ eMeasure (HQMF) implemented in EHRs
☒ other: Vital Statistics	☒ other: Vital Statistics

1.2. If an existing dataset was used, identify the specific dataset (*the dataset used for testing must be consistent with the measure specifications for target population and healthcare entities being measured; e.g., Medicare Part A claims, Medicaid claims, other commercial insurance, nursing home MDS, home health OASIS, clinical registry*).
CA- California Patient Discharge Data/Emergency Department/Ambulatory Surgery/Birth Cohort linked with vital statistics data.

1.3. What are the dates of the data used in testing? Discharges in 1995 through 2009 were used. There are no secular trends in these data to suggest that different years of data would result in substantively different hospital-level rates of these outcomes.

1.4. What levels of analysis were tested? (*testing must be provided for all the levels specified and intended for measure implementation, e.g., individual clinician, hospital, health plan*)

Measure Specified to Measure Performance of: (*must be consistent with levels entered in item S.26*)	**Measure Tested at Level of:**
☐ individual clinician	☐ individual clinician
☐ group/practice	☐ group/practice
☒ hospital/facility/agency	☒ hospital/facility/agency
☐ health plan	☐ health plan
☒ other: State	☒ other: State

1.5. How many and which measured entities were included in the testing and analysis (by level of analysis and data source)? (*identify the number and descriptive characteristics of measured entities included in the analysis (e.g., size, location, type); if a sample was used, describe how entities were selected for inclusion in the sample*)
For the CA data, 322,247 patients were identified from 357 hospitals. The hospital size varied across the dataset. Both teaching and non-teaching hospitals were included in testing.

Table 1: Characteristics of included hospitals

	Number of Patients	Number of Hospitals	Patients per Hospital	Hospitals at each NICU Level N (%)					
				1	2a	2b	3a	3b	3c
1995	20,476	305	67.13 (72.87)	191 (62.62%)	6 (1.97%)	12 (3.93%)	20 (6.56%)	66 (21.64%)	10 (3.28%)
1996	19,708	308	63.99 (67.86)	191 (62.01%)	8 (2.6%)	15 (4.87%)	18 (5.84%)	66 (21.43%)	10 (3.25%)
1997	19,286	312	61.81 (64.77)	194 (62.18%)	7 (2.24%)	14 (4.49%)	22 (7.05%)	67 (21.47%)	8 (2.56%)

1998	19,572	306	63.96 (69.61)	184 (60.13%)	6 (1.96%)	12 (3.92%)	22 (7.19%)	69 (22.55%)	13 (4.25%)
1999	18,782	301	62.4 (66.36)	182 (60.47%)	7 (2.33%)	14 (4.65%)	20 (6.64%)	66 (21.93%)	12 (3.99%)
2000	19,234	293	65.65 (68.32)	169 (57.68%)	9 (3.07%)	14 (4.78%)	23 (7.85%)	67 (22.87%)	11 (3.75%)
2001	18,867	285	66.2 (67.36)	161 (56.49%)	9 (3.16%)	14 (4.91%)	20 (7.02%)	70 (24.56%)	11 (3.86%)
2002	18,765	279	67.26 (68.31)	156 (55.91%)	8 (2.87%)	12 (4.3%)	22 (7.89%)	71 (25.45%)	10 (3.58%)
2003	19,689	283	69.57 (73.48)	162 (57.24%)	8 (2.83%)	12 (4.24%)	20 (7.07%)	70 (24.73%)	11 (3.89%)
2004	20,422	271	75.36 (79.95)	143 (52.77%)	8 (2.95%)	13 (4.8%)	25 (9.23%)	71 (26.2%)	11 (4.06%)
2005	21,664	267	81.14 (93.93)	142 (53.18%)	7 (2.62%)	15 (5.62%)	22 (8.24%)	73 (27.34%)	8 (3%)
2006	21,772	263	82.78 (97.18)	138 (52.47%)	7 (2.66%)	14 (5.32%)	21 (7.98%)	72 (27.38%)	11 (4.18%)
2007	21,660	258	83.95 (94.84)	135 (52.33%)	8 (3.1%)	14 (5.43%)	20 (7.75%)	73 (28.29%)	8 (3.1%)
2008	19,972	260	76.82 (87.86)	136 (52.31%)	9 (3.46%)	16 (6.15%)	22 (8.46%)	67 (25.77%)	10 (3.85%)
2009	18,583	252	73.74 (78.79)	132 (52.38%)	7 (2.78%)	14 (5.56%)	21 (8.33%)	67 (26.59%)	11 (4.37%)
Data Description: CA hospital data 1995-2009									

1.6. How many and which <u>patients</u> were included in the testing and analysis (by level of analysis and data source)?
(*identify the number and descriptive characteristics of patients included in the analysis (e.g., age, sex, race, diagnosis); if a sample was used, describe how patients were selected for inclusion in the sample*)

TABLE 2. CA ADMINISTRATIVE DATA	
INFANT DEMOGRAPHICS	
MALE	160815 (53.88%)
GESTATIONAL AGE (WEEKS)	
23	3819 (1.28%)
24	4794 (1.61%)
25	5841 (1.96%)
26	6803 (2.28%)
27	8378 (2.81%)
28	10376 (3.48%)
29	13267 (4.45%)
30	18709 (6.27%)
31	25856 (8.66%)
32	37835 (12.68%)
33	59768 (20.03%)
34	103006 (34.51%)
BROCHOPULMNARY DYSPLASIA (BPD)	10944 (3.67%)
NECROTIZING ENTEROCOLITIS (NEC)	4667 (1.56%)
RETINOPATHY OF PREMATURITY (ROP)	15670 (5.25%)
INTRAVENTRICULAR HEMORRHAGE (IVH)	12474 (4.18%)

READMISSION RATES	
7 DAYS	6621 (2.22%)
14 DAYS	9806 (3.29%)
30 DAYS	15682 (5.25%)
90 DAYS	29550 (9.9%)
365 DAYS	56828 (19.04%)
MATERNAL DEMOGRAPHICS N (%)	
AGE GROUP	
<18 YEARS	15666 (5.25%)
18-35 YEARS	235816 (79.02%)
>35 YEARS	46928 (15.73%)
EDUCATION	
NO HIGH SCHOOL	36958 (12.7%)
SOME HIGH SCHOOL	62753 (21.57%)
HIGH SCHOOL DIPLOMA/GED	82916 (28.49%)
AT LEAST SOME COLLEGE	108367 (37.24%)
RACE	
WHITE, NON-HISPANIC	93815 (31.43%)
BLACK, NON-HISPANIC	28816 (9.66%)
HISPANIC	138681 (46.47%)
ASIAN/PACIFIC ISLANDER	37140 (12.44%)
SOURCE OF PAYMENT	
FFS	9299 (3.23%)
HMO	124435 (43.21%)
FEDERAL	114754 (39.85%)
FEDERAL MC	33120 (11.5%)
OTHER MC	731 (0.25%)
OTHER	2499 (0.87%)
UNINSURED	3109 (1.08%)

1.7. If there are differences in the data or sample used for different aspects of testing (e.g., reliability, validity, exclusions, risk adjustment), identify how the data or sample are different for each aspect of testing reported below. Validity analyses of the CA data were limited to hospitals that discharged a minimum of 50 eligible patients per year to improve stability of the metrics.

1.8 What were the patient-level sociodemographic (SDS) variables that were available and analyzed in the data or sample used? For example, patient-reported data (e.g., income, education, language), proxy variables when SDS data are not collected from each patient (e.g. census tract), or patient community characteristics (e.g. percent vacant housing, crime rate).
In the CA data, source of payment are available at the patient level from birth certificates and hospital readmission data. Such information was categorized into different forms of private insurance (fee for service, health maintenance organization) and public insurance (Federal; Federal managed care). Other forms of insurance made up < 1% of the dataset. Uninsurance was found in approximately 1% of the data.

_____

2a2. RELIABILITY TESTING

Note*: If accuracy/correctness (validity) of data elements was empirically tested, separate reliability testing of data elements is not required – in 2a2.1 check critical data elements; in 2a2.2 enter "see section 2b2 for validity testing of data elements"; and skip 2a2.3 and 2a2.4.*

2a2.1. What level of reliability testing was conducted? (*may be one or both levels*)

☐ **Critical data elements used in the measure** (*e.g., inter-abstractor reliability; data element reliability must address ALL critical data elements*)

☒ **Performance measure score** (e.g., *signal-to-noise analysis*)

2a2.2. For each level checked above, describe the method of reliability testing and what it tests (*describe the steps—do not just name a method; what type of error does it test; what statistical analysis was used*)

Reproducibility of the results were calculated using Spearman-Brown statistics. Briefly, a 50% random sample of patients was drawn from each health care unit (hospital), and risk-adjustment models were calculated. Then, a second 50% sample was chosen and Spearman rank sum correlation coefficients were calculated. This metric assesses the influence of changes to the casemix of a hospital, where one assumes that the 50% sample provides an "alternative" insight into the measured readmission rates at each hospital or state.

2a2.3. For each level of testing checked above, what were the statistical results from reliability testing? (e.*g., percent agreement and kappa for the critical data elements; distribution of reliability statistics from a signal-to-noise analysis*)

Spearman-Brown correlation coefficients are presented by volume quartile because some of the smallest hospitals have few events to examine.

Readmission 30 Day	Spearman-Brown correction
Volume Lowest Quartile	0.736
Volume 2nd Quartile	0.931
Volume 3rd Quartile	0.880
Volume Highest Quartile	0.963

2a2.4 What is your interpretation of the results in terms of demonstrating reliability? (i.*e., what do the results mean and what are the norms for the test conducted?*)
Data show acceptable reliability, with Spearman-Brown split sample correlations above 0.70 for each volume strata. Stronger reliabilities above 0.85 are found in the three largest volume strata. Such reliability is similar to that suggested in other work from adult studies (re: Press MJ. Health Affairs 2013; 32: 1083-1091).

_____

2b2. VALIDITY TESTING

2b2.1. What level of validity testing was conducted? (*may be one or both levels*)

☐ **Critical data elements** (*data element validity must address ALL critical data elements*)

☒ **Performance measure score**

- ☒ **Empirical validity testing**
- ☐ **Systematic assessment of face validity of <u>performance measure score</u> as an indicator** of quality or resource use (*i.e., is an accurate reflection of performance on quality or resource use and can distinguish good from poor performance*)

2b2.2. For each level of testing checked above, describe the method of validity testing and what it tests (*describe the steps—do not just name a method; what was tested, e.g., accuracy of data elements compared to authoritative source, relationship to another measure as expected; what statistical analysis was used)*
We present 2 tests of the validity of the measure. First, we examine the correlation between 30-day readmission rates and hospital volume. Correlation with volume was performed based on previous work suggesting a volume-outcome association with other potential measures of NICU quality, such as mortality rates (Re: Phibbs C. NEJM 2007; Rogowski JAMA 2004), and thus higher volumes are a structural measure of neonatal intensive care. This work parallels other work in the literature that suggests that higher volume hospitals have improved outcomes, likely secondarily to seeing more patients and implementing processes of care to improve their outcomes. We hypothesize that there should be a larger association between hospital volume and 30-day readmission rates compared to complication rates.

Second, we examine other hypothesized measures of quality that may assess a different aspect of NICU quality. These structural and outcome measures include risk-adjusted rates of common complications of premature birth where variation is known, such as bronchopulmonary dysplasia (BPD), intraventricular hemorrhage (IVH), necrotizing enterocolitis (NEC), and retinopathy of prematurity (ROP). Complication rates have been suggested by the Institute of

Medicine as appropriate surrogates of hospital quality. These complications have also been demonstrated to be important risk factors for the development of long-term neurodevelopmental delay and cerebral palsy (Schmidt B. JAMA. 2003: 289(9), 1124-1129). However, the processes of care that may reduce the development of complications throughout the hospital stay (improved respiratory ventilator management, improved feeding programs, hand hygiene programs) may only be modestly associated with processes of care that improve readmission rates (such as improved transitions of care, improved education of families, etc). These intermediate process measures are not available in any large scale population-based dataset. If we find an association between 30-day readmission rates and neonatal complications, it would call into question whether the extra time needed to quantify readmission rates should be undertaken by hospitals, insurers, state agencies, and other bodies interested in assessing the quality of neonatal intensive care. For further information on this topic, see Lorch JAMA Pediatrics 2014.

2b2.3. What were the statistical results from validity testing? (*e.g., correlation; t-test*)

Table 3. Hospital Volume Correlation	
Quintile	30-day risk adjusted readmission rate
Lowest	0.062
2	0.057
3	0.053
4	0.051
Highest	0.052
P-value, test of trend	0.032

All complication rates reported in the following tables are risk adjusted using same model as readmission rates. Spearman correlation coefficients are presented; similar results were found with Pearson's correlation coefficients.

Table 4. Correlation of Adjusted Readmission Rate with Complication Rates, Spearman correlation coefficients	
BPD	0.12
ROP	0.14
NEC	0.12
IVH	0.16

All correlations had a p-value > 0.05.

2b2.4. What is your interpretation of the results in terms of demonstrating validity? (i.*e., what do the results mean and what are the norms for the test conducted?*)
For the California data, with a greater number of hospitals, we found a volume-outcome association with rates of readmission within 30 days after discharge: lower volume hospitals had higher rates of readmissions through this time period. While the absolute difference was 1.0% between the smallest and largest quintile, this difference was a 16% relative increase in readmission rates.

As we hypothesized, there were poor to no correlations between hospital-level risk adjusted rates of complications and risk-adjusted readmission rates. Such lack of correlations is similar to that found in most studies of adult readmission rates, which hypothesize that readmission rates assess a different aspect of the quality of NICUs compared with that assessed by complication rates.

2b3. EXCLUSIONS ANALYSIS

NA ☐ no exclusions — *skip to section 2b4*

2b3.1. Describe the method of testing exclusions and what it tests (*describe the steps—do not just name a method; what was tested, e.g., whether exclusions affect overall performance scores; what statistical analysis was used*)

Two groups of patients were excluded from this measure:

1) Infants with congenital anomalies as described in S10 (codes are provided in the attached Data Dictionary).
2) Hospitals with fewer than 50 observations per time period.

Exclusion #1 occurred because of potential loss to face validity with a measure where infants with known anomalies that will require further hospitalizations were included in a measure. These infants do not receive care randomly throughout the health care system, and there is known concern on the part of hospitals that care for these infants that a readmission measure may be unfair to these hospitals when comparing readmission rates to hospitals who do not typically care for these infants. Such exclusions are common in neonatal intensive care measures.

Exclusion #2 occurred because of difficulties in measuring hospitals with small denominators. Prior work suggests that 50 observations are needed to produce a reliable measure, in the sense that many of these hospitals have no observed outcomes, or by having only 1-2 additional observed outcomes, the observed/expected rate of the outcome at that hospital changes substantially.

2b3.2. What were the statistical results from testing exclusions? (*include overall number and percentage of individuals excluded, frequency distribution of exclusions across measured entities, and impact on performance measure scores*)

Table 5. Description of CA Patients and Hospitals Excluded

	Patients Excluded for Congenital Anomalies	Hospitals Excluded for small volume, 1 year
1995-2009	20,648 (6.47%)	90 (0.03%)
1995	1,201 (5.54%)	5 (0.02%)
1996	1,234 (5.89%)	17 (0.08%)
1997	1,234 (6.01%)	4 (0.02%)
1998	1,201 (5.78%)	4 (0.02%)
1999	1,257 (6.27%)	10 (0.05%)
2000	1,233 (6.02%)	11 (0.05%)
2001	1,309 (6.49%)	4 (0.02%)
2002	1,331 (6.62%)	8 (0.04%)
2003	1,287 (6.13%)	10 (0.05%)
2004	1,402 (6.42%)	4 (0.02%)
2005	1,475 (6.37%)	5 (0.02%)
2006	1,598 (6.84%)	0 (0.0%)
2007	1,665 (7.14%)	5 (0.02%)
2008	1,610 (7.46%)	1 (0%)
2009	1,611 (7.98%)	2 (0.01%)

Rates of readmissions were substantively higher in the congenital anomaly group, with a mean increase of 9% higher.

2b3.3. What is your interpretation of the results in terms of demonstrating that exclusions are needed to prevent unfair distortion of performance results? (*i.e., the value outweighs the burden of increased data collection and analysis. Note: **If patient preference is an exclusion**, the measure must be specified so that the effect on the performance score is transparent, e.g., scores with and without exclusion*)

Exclusions for congenital anomalies is necessary to maintain face validity of the data, as these infants have higher risk of readmission compared to other infants that may not be risk adjusted in the analysis. For small hospitals,

there is a lack of reliability in intra-year data secondary to the small numbers of patients at a given hospital for a specified time period. This results in excess noise in the measure, limiting the ability to generalize results from these hospitals.

_____

2b4. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES

If not an intermediate or health outcome, or PRO-PM, or resource use measure, skip to section 2b5.

2b4.1. What method of controlling for differences in case mix is used?

☐ **No risk adjustment or stratification**

☒ **Statistical risk model with** Click here to enter number of factors **risk factors**

☐ **Stratification by** Click here to enter number of categories **risk categories**

☐ **Other,** Click here to enter description

2b4.2. If an outcome or resource use measure is not risk adjusted or stratified, provide rationale and analyses to demonstrate that controlling for differences in patient characteristics (case mix) is not needed to achieve fair comparisons across measured entities.

N/A

2b4.3. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors (clinical factors or sociodemographic factors) used in the statistical risk model or for stratification by risk (*e.g., potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of p<0.10; correlation of x or higher; patient factors should be present at the start of care*)

Risk factors were chosen based on a review of the literature using population-based datasets available to public agencies. All factors are available at the time the infant is delivered, i.e. the start of care for the infant.

Medical factors such as gestational age was included based on prior work that suggests greater risk of readmission as an infant is more premature (Re: Ray K, Hospital Pediatrics, 2013). This increase in risk occurs independent of the development of other complications of preterm birth that are also known risk factors for readmission, such as bronchopulmonary dysplasia (higher rate of readmission with exposure to viral illness; higher risk of asthma; higher rate of failure-to-thrive); necrotizing enterocolitis (higher rate of readmission with exposure to viral illness); and intraventricular hemorrhage (higher rate of readmission secondary to hydrocephalus and VP-shunt malfunction). Sociodemographic data were included because of prior data from our group and others (Lorch Academic Pediatrics 2014; Ray Hospital Pediatrics 2013; Ray Academic Pediatrics 2010) that have found an association between readmission and maternal age, likely secondary to differences in provider practice related to a provider's assessment of the compliance of the family and the parental preferences toward a hospital admission; and higher rates of readmission in publicly-insured children. We included race/ethnicity in the models for completeness sake and because there were identified differences in these data.

2b4.4a. What were the statistical results of the analyses used to select risk factors?

Exemplar risk adjustment models are shown below for the 30 day readmission rate.

Table 16. Risk Estimate Model				
	OR	**OR LCL**	**OR UCL**	**P-Value**
Intercept				<.0001
Gestational Age (Reference Group: 34 Weeks)				
23	0.762	0.646	0.897	0.0011
24	1.112	0.982	1.259	0.0956
25	1.11	0.991	1.243	0.07
26	1.17	1.054	1.298	0.0032
27	1.12	1.017	1.234	0.0213
28	1.098	1.005	1.201	0.0392
29	1.085	1.001	1.177	0.0465
30	1.017	0.947	1.093	0.6387

31	1	0.938	1.066	0.9968
32	0.962	0.91	1.018	0.1791
33	0.946	0.902	0.993	0.025
Male	1.177	1.138	1.217	<.0001
BPD	1.34	1.234	1.455	<.0001
ROP	1.443	1.343	1.551	<.0001
NEC	1.52	1.368	1.688	<.0001
IVH Any	1.131	1.047	1.222	0.0018
Maternal Race/Ethnicity (Reference Group: White, Non-Hispanic)				
Black	0.937	0.88	0.998	0.0446
Hispanic	1.063	1.018	1.11	0.0052
Asian/ Pacific Islander	0.962	0.908	1.02	0.193
Maternal Age Group (Reference Group: 18-35 Years)				
< 18	0.966	0.894	1.043	0.3779
> 35	0.928	0.884	0.975	0.0028
Source of Payment (Reference Group: FFS)				
HMO	1.019	0.919	1.131	0.7196
Federal	1.346	1.211	1.496	<.0001
Federal MC	1.671	1.496	1.866	<.0001
Other MC	1.194	0.844	1.688	0.3164
Other	0.825	0.653	1.043	0.1085
Uninsured	0.629	0.495	0.800	0.0002

2b4.4b. Describe the analyses and interpretation resulting in the decision to select SDS factors (e.g. prevalence of the factor across measured entities, empirical association with the outcome, contribution of unique variation in the outcome, assessment of between-unit effects and within-unit effects)
SDS factors were included based on their prevalence in the dataset and their association with the outcome as noted in the above table. Much of the background is described above. The resulting results support the inclusion of these factors, as there were associations between 30-day readmission and insurance status (higher rates in publicly insured, consistent with prior work); maternal age (lower rates in older women, possibly secondary to the types of providers treating these infants or differences in the patient preferences).

2b4.5. Describe the method of testing/analysis used to develop and validate the adequacy of the statistical model or stratification approach (*describe the steps—do not just name a method; what statistical analysis was used*)
Model discrimination was measured using c-statistics, while model calibration was assessed using the Hosmer-Lemeshow test with 10 equal risk groups.
Provide the statistical results from testing the approach to controlling for differences in patient characteristics (case mix) below.
If stratified, skip to 2b4.9
2b4.6. Statistical Risk Model Discrimination Statistics (*e.g., c-statistic, R-squared*)**:**
The c-statistic result for 30-day readmission was found to be 0.58

2b4.7. Statistical Risk Model Calibration Statistics (*e.g., Hosmer-Lemeshow statistic*) and **2b4.8. Statistical Risk Model Calibration – Risk decile plots or calibration curves**

Table 18. Hosmer-Lemeshow Statistic Results

outcome	Group	Number of events	Predicted number of events
30-day rate (p=0.29)	1	983	1015.34
	2	1118	1117.32

3	1196	1209.34
4	1144	1162.93
5	1338	1336.15
6	1454	1433.87
7	1516	1539.94
8	1640	1680.86
9	1917	1813.27
10	2531	2528.28

2b4.9. Results of Risk Stratification Analysis:

N/A

2b4.10. What is your interpretation of the results in terms of demonstrating adequacy of controlling for differences in patient characteristics (case mix)? (i*.e., what do the results mean and what are the norms for the test conducted*)
The risk adjustment model shows modest discrimination and strong calibration. While the c-statistic for the risk adjustment model is less than the desired level of 0.7, this model for predicting readmission risk using medical factors is similar to prior work from our group and others, with a c-statistic between 0.6 and 0.65, and in other outcome measures such as length of stay and health care costs (Ray KN. Hospital Pediatrics. 2013;3:194-203; Lorch SA. Hlth Serv Res. 2010;45:24-41.; Schwartz, M., and A. S. Ash. 2003. ''Evaluating Risk-Adjustment Model Empirically.'' In Risk Adjustment for Measuring Health Care Outcomes, edited by L. I. Iezzoni. Chicago: Health Administration Press). Secondly, as we have discussed in other portions of this document, we initially considered a non-risk adjusted metric. This unadjusted rate argues that any readmission is undesired – although a zero percent readmission rate is not practical to achieve and may have other untoward consequences such as longer initial lengths of stay. However, to improve face validity, we included factors that have been associated with higher readmission rates in the literature. Of note, the average percent change between the unadjusted and risk-adjusted rates was 0.75%, with a range of -3.4% to 5.6%. Thus, the inclusion of risk adjustment was felt to be necessary by clinicians to improve the face validity of the results, but made little difference to the actual measurements.

These factors, though, would not account for the variation in risk the result from hospital-level differences in the processes of discharge that we would like to measure, such as parental education and processes to ensure the transfer of care from the inpatient to the outpatient setting. These results also suggest that medical factors cannot totally predict the risk of readmission in these infants, supporting the idea that readmissions measure a construct separate from medical complication rates.

2b4.11. Optional Additional Testing for Risk Adjustment (*not required, but would provide additional support of adequacy of risk model, e.g., testing of risk model in another data set; sensitivity analysis for missing data; other methods that were assessed*)
N/A

_____

2b5. IDENTIFICATION OF STATISTICALLY SIGNIFICANT & MEANINGFUL DIFFERENCES IN PERFORMANCE
2b5.1. Describe the method for determining if statistically significant and clinically/practically meaningful differences in performance measure scores among the measured entities can be identified (*describe the steps—do not just name a method; what statistical analysis was used? Do not just repeat the information provided related to performance gap in 1b)*

Outliers were defined as those entities with rates greater than 75th percentile value + 1.5*interquartile range (aka "upper adjacent value"), or less than 25th percentile - 1.5*interquartile range (aka "lower adjacent value").

2b5.2. What were the statistical results from testing the ability to identify statistically significant and/or clinically/practically meaningful differences in performance measure scores across measured entities? (e.g., *number*

and percentage of entities with scores that were statistically significantly different from mean or some benchmark, different from expected; how was meaningful difference defined)
Statistics for unadjusted readmission rates among California hospitals:

Table. Readmission within 30 days, unadjusted									
					Deciles				
Mean	Std Dev	Min	Max	IQR	10%	30%	50%	70%	90%
0.058	0.055	0.000	0.500	(0.025-0.078)	0.000	0.029	0.045	0.069	0.116

Table. Readmission within 30 days, risk adjusted									
					Deciles				
Mean	Std Dev	Min	Max	IQR	10%	30%	50%	70%	90%
0.060	0.056	0.000	0.528	(0.025-0.079)	0.000	0.030	0.048	0.071	0.115

Outlier hospitals: All outlier hospitals had rates above the upper adjusted rate. In each case, approximately 8% of the hospital population met the definition described above.

2b5.3. What is your interpretation of the results in terms of demonstrating the ability to identify statistically significant and/or clinically/practically meaningful differences in performance across measured entities? (i.*e., what do the results mean in terms of statistical and meaningful differences?*)
There are outlier hospitals identified by this measure.

2b6. COMPARABILITY OF PERFORMANCE SCORES WHEN MORE THAN ONE SET OF SPECIFICATIONS
***If only one set of specifications, this section can be skipped*.**
Note*: This item is directed to measures that are risk-adjusted (with or without SDS factors)* ***OR*** *to measures with more than one set of specifications/instructions (e.g., one set of specifications for how to identify and compute the measure from medical record abstraction and a different set of specifications for claims or eMeasures). It does not apply to measures that use more than one source of data in one set of specifications/instructions (e.g., claims data to identify the denominator and medical record abstraction for the numerator).* ***Comparability is not required when comparing performance scores with and without SDS factors in the risk adjustment model. However, if comparability is not demonstrated for measures with more than one set of specifications/instructions, the different specifications (e.g., for medical records vs. claims) should be submitted as separate measures.***

2b6.1. Describe the method of testing conducted to compare performance scores for the same entities across the different data sources/specifications (*describe the steps—do not just name a method; what statistical analysis was used*)
N/A
2b6.2. What were the statistical results from testing comparability of performance scores for the same entities when using different data sources/specifications? (*e.g., correlation, rank order*)
N/A
2b6.3. What is your interpretation of the results in terms of the differences in performance measure scores for the same entities across the different data sources/specifications? (i.*e., what do the results mean and what are the norms for the test conducted*)
N/A

2b7. MISSING DATA ANALYSIS AND MINIMIZING BIAS
2b7.1. Describe the method of testing conducted to identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias (*describe the*

steps—do not just name a method; what statistical analysis was used)
We compared hospital rates of readmissions including patients with a documented gestational age (provided throughout the document), and hospital rates after all patients were included when gestational age was multiply imputed 5 times using birth weight, length of stay, maternal comorbid conditions, and the presence of complications of preterm birth including BPD, IVH, NEC, and ROP. Each of these factors is associated with gestational age, with birth weight having the strongest correlation. We included maternal comorbid conditions to help assist in the identification of infants "small-for-gestational age" (i.e. whose birth weight was < 10^{th} percentile for gestational age based on national growth curves) or "large-for gestational age (i.e. whose birth weight was > 90^{th} percentile for gestational age), as these infants would potentially be misclassified by use of birth weight alone. Similar information to refine the gestational age estimate was found by inclusion of length of stay (greater in infants of lower gestational age) and complications of preterm birth (more common in infants of lower gestational age). Multiple imputation techniques allow for reduction of bias secondary to the systemtic exclusion of infants missing gestational age, such as infants who die in the delivery room. Multiple imputation, though, allows for the models to properly account for the fact that the estimated gestational age is only an estimate and thus has a potential range of values – including just the best estimate from the model ignores this finding. After multiply imputing the gestational age, models are rerun. The results of the models are then combined using stratified data analysis techniques to arrive at a result. Comparisons of the models that include missing patients using these more complex statistical techniques to the reported models where these infants are excluded use Spearman rank correlation coefficient statistics.

2b7.2. What is the overall frequency of missing data, the distribution of missing data across providers, and the results from testing related to missing data? (*e.g., results of sensitivity analysis of the effect of various rules for missing data/nonresponse; if no empirical sensitivity analysis, identify the approaches for handling missing data that were considered and pros and cons of each*)
CA patients with a gestational age had a higher rate of readmission compared to those who did not, likely because patients missing GA tended to be of older gestational age using multiple imputation techniques. However, Spearman rank sum correlation coefficients between rates where infants with missing gestational age were excluded to rates where the GA for these infants was estimated using multiple imputation techniques were very high, at 0.82.

2b7.3. What is your interpretation of the results in terms of demonstrating that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias**?** (i.*e., what do the results mean in terms of supporting the selected approach for missing data and what are the norms for the test conducted; if no empirical analysis, provide rationale for the selected approach for missing data*)
Overall, patients with gestational age values for those metrics without vital statistics may overestimate the readmission rate for a given health care unit, but these values do not substantively change the rank ordering of the units. To reduce the burden of calculating these gestational ages in missing patients, we have proposed to omit infants without a documented gestational age. Assessment of the correlation in reported readmission rates between the proposed method (exclusion of infants without a gestational age) and the multiply imputed method should be performed for a specific dataset.

3. Feasibility

Extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

3a. Byproduct of Care Processes

For clinical measures, the required data elements are routinely generated and used during care delivery (e.g., blood pressure, lab test, diagnosis, medication order).

3a.1. **Data Elements Generated as Byproduct of Care Processes.**

Coded by someone other than person obtaining original information (e.g., DRG, ICD-9 codes on claims)

If other:

3b. Electronic Sources

The required data elements are available in electronic health records or other electronic sources. If the required data are not in electronic health records or existing electronic sources, a credible, near-term path to electronic collection is specified.

3b.1. **To what extent are the specified data elements available electronically in defined fields?** *(i.e., data elements that are needed to compute the performance measure score are in defined, computer-readable fields)*

ALL data elements are in defined fields in a combination of electronic sources

3b.2. **If ALL the data elements needed to compute the performance measure score are not from electronic sources, specify a credible, near-term path to electronic capture, OR provide a rationale for using other than electronic sources.**

3b.3. **If this is an eMeasure, provide a summary of the feasibility assessment in an attached file or make available at a measure-specific URL.**

Attachment:

3c. Data Collection Strategy

Demonstration that the data collection strategy (e.g., source, timing, frequency, sampling, patient confidentiality, costs associated with fees/licensing of proprietary measures) can be implemented (e.g., already in operational use, or testing demonstrates that it is ready to put into operational use). For eMeasures, a feasibility assessment addresses the data elements and measure logic and demonstrates the eMeasure can be implemented or feasibility concerns can be adequately addressed.

3c.1. **Describe what you have learned/modified as a result of testing and/or operational use of the measure regarding data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, time and cost of data collection, other feasibility/implementation issues.**

IF a PRO-PM, consider implications for both individuals providing PROM data (patients, service recipients, respondents) and those whose performance is being measured.

Our tests of the measure show a high degree of variation across hospitals, even after our attempts to adjust for differences in NICU case-mix. However, that implementation may be difficult due to missing data from administrative datasets in use at the state and federal level. Important adjusting variables such as gestational age and birth weight are not currently recorded consistently in MAX or like datasets, and thus accurate implementation of this metric will require new data collection, linkage with birth certificates, or more widespread and standardized use of EHR for publicly reported measures. We did find that including only patients with gestational age, versus using all patients with multiple imputation rates, did not appreciably change the rank ordering of hospitals. Whether similar results would be seen at the national level with more health care entities such as hospitals or outpatient providers outside of California is not clear. MAX data are only now capable of identifying hospitals or outpatient providers reliably, requiring data from 2009 or later to be linked to National Provider ID hospital numbers.

More generally, data quality and completeness in MAX varies by state. Some states do not report enrollment data, and none report claims for their state funded (S-CHIP) programs. Similarly, managed care claims are sometimes absent from MAX, and reported managed care data is not always validated before inclusion. We also historically found that inpatient records were missing in up to 20% of the identified claims in some states. We posit that this may occur because inpatient hospital administrative records that provide this information may not be linked or reported to MAX if the hospitalization occurs out-of-state. State-level rates of readmissions do not require these records, since CPT billing codes serve as a valid alternative and are present regardless of where the admission occurred. However, to construct hospital-level readmission rates, these records will become important to provide a complete picture of the variation in readmission rates across the hospitals that service patients of a given state Medicaid office. Another important feature is the need for all-hospital readmission rates, not just readmissions to the same hospital as the

discharging neonatal intensive care unit. Because NICUs are widely available in most health care systems, frequently hospitals that provide maternity care and care for infants requiring NICU care may not have a pediatrics floor. Thus, many readmissions occur to inpatient units that differ from the hospital where the infant was discharged from the NICU. Such information must cross not only hospitals, but also state lines for patients living near a state border. Admission records must then collect information about readmissions to all hospitals. Insurance records will contain this information, as well population-based data. Data from individual hospitals or consortiums must develop systems to track patients to other hospitals to ensure a valid measure.
An additional complication with the NICU Readmission measure, like any metric based on readmissions, is that it is very difficult to identify preventable readmissions from those which are necessary. There has not yet been a determination of the "optimal" level of readmissions in a state or hospital, so we cannot necessarily suggest that the lowest or highest observed rates are ideal, or where they fall relative to what we "should" observe. Many established quality metrics, including those of the CHIPRA Initial Core Set, strive for a 0% or 100% performance rate. Identification of a baseline number of expected events is a much more difficult prospect, and thus complicates the identification of outliers or underperformers. Additionally, we currently do not know what factors underlie the variation in readmission rates. While some of the variation could be related to the quality of care provided during the inpatient stay or discharge process, some might also be related to outpatient care quality or a child's access to services. Lastly, some variation due to severity may persist even after risk adjustment.

3c.2. **Describe any fees, licensing, or other requirements to use any aspect of the measure as specified** *(e.g., value/code set, risk model, programming code, algorithm)*.
N/A

4. Usability and Use

Extent to which potential audiences (e.g., consumers, purchasers, providers, policy makers) are using or could use performance results for both accountability and performance improvement to achieve the goal of high-quality, efficient healthcare for individuals or populations.

4a. Accountability and Transparency

Performance results are used in at least one accountability application within three years after initial endorsement and are publicly reported within six years after initial endorsement (or the data on performance results are available). If not in use at the time of initial endorsement, then a credible plan for implementation within the specified timeframes is provided.

4.1. Current and Planned Use

NQF-endorsed measures are expected to be used in at least one accountability application within 3 years and publicly reported within 6 years of initial endorsement in addition to performance improvement.

Planned	Current Use (for current use provide URL)
Public Reporting Quality Improvement with Benchmarking (external benchmarking to multiple organizations) Quality Improvement (Internal to the specific organization) Not in use	

4a.1. For each CURRENT use, checked above, provide:

- Name of program and sponsor
- Purpose
- Geographic area and number and percentage of accountable entities and patients included

4a.2. If not currently publicly reported OR used in at least one other accountability application (e.g., payment program, certification, licensing) what are the reasons? (*e.g., Do policies or actions of the developer/steward or accountable entities restrict access to performance results or impede implementation?*)

This metric was developed in part with the CHIPRA Pediatric Quality Measure Program funded by AHRQ. It will be part of a collection of measures publicized as a resource for policymakers, advocates, consumers, and purchasers of children's health care; researchers; providers; and others interested in identifying valid, reliable, and feasible quality measures of children's health care. As such, it is one of many measures included in a portfolio for improved measures.

4a.3. If not currently publicly reported OR used in at least one other accountability application, provide a credible plan for implementation within the expected timeframes -- any accountability application within 3 years and publicly reported within 6 years of initial endorsement. (*Credible plan includes the specific program, purpose, intended audience, and timeline for implementing the measure within the specified timeframes. A plan for accountability applications addresses mechanisms for data aggregation and reporting.*)

This metric will be included in the portfolio of measures publicized by the CHIPRA PQMP program, available to all for use.

4b. Improvement

Progress toward achieving the goal of high-quality, efficient healthcare for individuals or populations is demonstrated. If not in use for performance improvement at the time of initial endorsement, then a credible rationale describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.

4b.1. Progress on Improvement. (Not required for initial endorsement unless available.)

Performance results on this measure (current and over time) should be provided in 1b.2 and 1b.4. Discuss:

- **Progress (trends in performance results, number and percentage of people receiving high-quality healthcare)**
- **Geographic area and number and percentage of accountable entities and patients included**

4b.2. **If no improvement was demonstrated, what are the reasons? If not in use for performance improvement at the time of initial endorsement, provide a credible rationale that describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.**

4c. Unintended Consequences

The benefits of the performance measure in facilitating progress toward achieving high-quality, efficient healthcare for individuals or populations outweigh evidence of unintended negative consequences to individuals or populations (if such evidence exists).

4c.1. **Were any unintended negative consequences to individuals or populations identified during testing; OR has evidence of unintended negative consequences to individuals or populations been reported since implementation? If so, identify the negative unintended consequences and describe how benefits outweigh them or actions taken to mitigate them.**

There have been no unintended negative consequences identified.

5. Comparison to Related or Competing Measures

If a measure meets the above criteria and there are endorsed or new related measures (either the same measure focus or the same target population) or competing measures (both the same measure focus and the same target population), the measures are compared to address harmonization and/or selection of the best measure.

5. Relation to Other NQF-endorsed Measures
Are there related measures (conceptually, either same measure focus or target population) or competing measures (conceptually both the same measure focus and same target population)? If yes, list the NQF # and title of all related and/or competing measures.
Yes

5.1a. **List of related or competing measures (selected from NQF-endorsed measures)**
2393 : Pediatric All-Condition Readmission Measure

5.1b. **If related or competing measures are not NQF endorsed please indicate measure title and steward.**
The target population of 2393 Pediatric All-Condition Readmission Measure differs from this measure. This measure is specific to the Neonatal population.

5a. Harmonization

The measure specifications are harmonized with related measures;
OR
The differences in specifications are justified

5a.1. **If this measure conceptually addresses EITHER the same measure focus OR the same target population as NQF-endorsed measure(s):**
Are the measure specifications completely harmonized?
No

5a.2. **If the measure specifications are not completely harmonized, identify the differences, rationale, and impact on interpretability and data collection burden.**
This measure is specific to the Neonatal population.

5b. Competing Measures

The measure is superior to competing measures (e.g., is a more valid or efficient way to measure);
OR
Multiple measures are justified.

5b.1. **If this measure conceptually addresses both the same measure focus and the same target population as NQF-endorsed measure(s):**
Describe why this measure is superior to competing measures (e.g., a more valid or efficient way to measure quality); OR provide a rationale for the additive value of endorsing an additional measure. (Provide analyses when possible.)
This measure provides the additive value of being specific to the unique patient population seen in the NICU.

Appendix

A.1 Supplemental materials may be provided in an appendix. All supplemental materials (such as data collection instrument or methodology reports) should be organized in one file with a table of contents or bookmarks. If material pertains to a specific submission form number, that should be indicated. Requested information should be provided in the submission form and required attachments. There is no guarantee that supplemental materials will be reviewed.
Attachment **Attachment:** Appendix.docx

Contact Information

Co.1 **Measure Steward (Intellectual Property Owner):** The Children´s Hospital of Philadelphia
Co.2 **Point of Contact:** Scott, Lorch, lorch@email.chop.edu, 215-590-1714-
Co.3 **Measure Developer if different from Measure Steward:** The Children´s Hospital of Philadelphia
Co.4 **Point of Contact:** Scott, Lorch, lorch@email.chop.edu, 215-590-1714-

Additional Information

Ad.1 **Workgroup/Expert Panel involved in measure development**
Provide a list of sponsoring organizations and workgroup/panel members' names and organizations. Describe the members' role in measure development.

Measure Developer/Steward Updates and Ongoing Maintenance
Ad.2 **Year the measure was first released:** 2016
Ad.3 **Month and Year of most recent revision:** 02, 2016
Ad.4 **What is your frequency for review/update of this measure?**
Ad.5 **When is the next scheduled review/update for this measure?**

Ad.6 **Copyright statement:**
Ad.7 **Disclaimers:**

Ad.8 **Additional Information/Comments:**

MEASURE WORKSHEET

This document summarizes the evaluation of the measure as it progresses through NQF's Consensus Development Process (CDP). The information submitted by measure developers/stewards is included after the Brief Measure Information, Preliminary Analysis, and Pre-meeting Public and Member Comments sections.

To navigate the links in the worksheet: Ctrl + click link to go to the link; ALT + LEFT ARROW to return

Brief Measure Information

NQF #: 2895
Measure Title: Thermal Condition of Low Birthweight Neonates Admitted to Level 2 or Higher Nurseries in the First 24 Hours of Life: A PQMP Measure
Measure Steward: University Hospitals Cleveland Medical Center
Brief Description of Measure: This measure describes in terms of admission temperature the status of live-born neonates less than 2,500 grams that are admitted to a Level 2 or higher nursery.

This measure reports on the temperature at admission. Temperatures are reported both in categorical terms and as a distribution. The distribution should be presented as a cumulative incidence curve with a chart to present key moments in the distribution. The categorization data may be presented in chart or graphical form, such as a pie chart, with parents. Each admission is categorized into one of five strata on the basis of their admission temperature. The strata, which were defined by our expert panel, are cold (<34.5), very cool (34.51-35.50), cool (35.51-36.50), about right (36.51-37.50) and overly warm (>37.5). All temperatures are analyzed using degrees Celsius and reported to one decimal place. The FIRST temperature taken in the nursery is to be recorded and used.

To avoid the potential for gaming the measure by delaying a recorded temperature after arrival, the results are stratified in three ways:

- Main Stratum: Time between arrival at Level 2 or higher nursery is between 0 and 15 minutes.
- Delayed stratum: Time between arrival at Level 2 or higher nursery is more than 15 minutes.
- Other: Inadequate documentation to determine timing of temperature

Developer Rationale: Inpatient perinatal care was assigned to CAPQuaM as a PQMP priority by the Agency for Healthcare Research and Quality with the active consultation of the Centers for Medicare & Medicaid Services. After initial assignment, conversations between CAPQuaM, AHRQ, and CMS resulted in a decision for CAPQuaM to undertake the development of measures related to the temperature of low birthweight neonates. We developed this measure in close consultation with our Consortium partners at the New York State Department of Health, including the Office of Health Insurance Program/New York State Medicaid and the Division of Family Health.

Hypothermia in neonates has been identified as associated with poor outcomes since the 19th century and is a firmly established construct. We note that there is evidence that management can enhance thermal outcomes. In work that preceded the CAPQuaM work, our team demonstrated that the temperature of low birthweight neonates is variable, and is highly consequential in terms of critical outcomes like survival and intraventricular hemorrhage. Institutional anecdotal evidence supports literature observations that thermal management can be managed and improved at the unit level with improved outcomes.

Numerator Statement: The metric of interest is the temperature upon arrival to the Level 2 or higher nursery that is being assessed. This measure does not have the form of numerator and denominator. It is a distribution. We ask for reporting of the distribution in terms of five categories across the distribution, in terms of key moments in the distribution, and as a graphical presentation of the distribution. This is an information rich measure. Accountability entities may choose to use any of various components for their emphasis (alone or in combination), including percent

"about right", mean or median temperatures, or value of the 10th or 25th percentiles, and the inter-percentile range.

There is an eligible population of newborns, which could be considered the denominator.
In lieu of a numerator, this measure reports the distribution of temperatures, using both numbers and a graph. In order to allow for reporting of key factors of interest to the accountability entity, this measure is specified to report that distribution in a variety of ways. This measure offers users (the accountability entity) the option to focus on one or more key substantive aspects of thermal outcomes in the defined population.

Data Elements:

-- Temperature to first decimal place
-- Units of temperature (Celsius, Fahrenheit). Those measured in Fahrenheit should be converted to Celsius. Celsius=(Fahrenheit less 32) times 5 divided by 9.
-- Time that temperature was measured
-- Time of arrival to the nursery (not time that admission was done)

State and County of residence OR zip code of mother
-- Optional: Method of temperature measurement (axillary, rectal, skin, tympanic)
Denominator Statement: All newborn infants born in a medical facility with birthweights less than 2,500 grams and admitted to a level 2 or higher nursery within 24 hours of life, other than those excluded.
Denominator Exclusions: Neonates with anencephaly, who receive only comfort care in the Level 2 or higher nursery, or those who die or are placed intentionally on a pre-existing hypothermia protocol prior to the 15 minute after arrival specification time.

Measure Type: Outcome
Data Source: Administrative claims, Electronic Clinical Data : Electronic Health Record, Other, Paper Medical Records
Level of Analysis: Facility, Health Plan, Integrated Delivery System, Population : Community, Population : County or City, Population : Regional, Population : State

IF Endorsement Maintenance – Original Endorsement Date: Most Recent Endorsement Date:

New Measure -- Preliminary Analysis

Criteria 1: Importance to Measure and Report

1a. Evidence

1a. Evidence. The evidence requirements for a *process or intermediate outcome* measure is that it is based on a systematic review (SR) and grading of the body of empirical evidence where the specific focus of the evidence matches what is being measured.

The developer provides the following evidence for this measure:

- **Systematic Review of the evidence specific to this measure?** ☒ **Yes** ☐ **No**
- **Quality, Quantity and Consistency of evidence provided?** ☒ **Yes** ☐ **No**
- **Evidence graded?** ☐ **Yes** ☒ **No**

Evidence Summary

- This measure is an intermediate clinical outcome.
- A conceptual model is provided describing the links between neonatal temperature and health outcomes.
- Guidelines from WHO indicate that "thermal protection of the newborn is a set of continuing measures, which starts at birth, to ensure that he maintains a body temperature of 36.5°C to 37.5°C."
- Two Cochrane systematic reviews are described that relate hypothermia to neonatal death and describe a variety of effective methods to keep a preterm baby warm. QQC is provided for one of the SRs.
- The developer provides results from their own work relating the findings of neonatal temperature and neonatal

mortality:

Impact of Temperature on Mortality (Kleinman, Howell et al, not yet published)	
Change Temperature (Celsius)	Net Benefit to Mortality*
From 34 to 35	2.8%
From 35 to 36	2.4%
From 36 to 37	2.0%
From 34 to 36	5.2%
From 35 to 37	4.4%
From 34 to 37	7.2%
*All mortality differences are derived from an adjustment model that includes clincial and demographic factors including race/ethnicity and 5 minute Apgar score.	

Exception to evidence NA

Guidance from the Evidence Algorithm

Not a health outcome measure (Box 1) →intermediate clinical outcome (Box 3) → SR with QQC → high

Questions for the Committee:

- *What is the relationship of this measure to patient outcomes?*
- *How strong is the evidence for this relationship?*
- *Is the evidence directly applicable to the process of care being measured?*

Preliminary rating for evidence: ☒ **High** ☐ **Moderate** ☐ **Low** ☐ **Insufficient**

1b. Gap in Care/Opportunity for Improvement and 1b. Disparities

1b. Performance Gap. The performance gap requirements include demonstrating quality problems and opportunity for improvement.

The developer provides data from testing the measure:

- In our study of 7,553 neonates admitted to Level 2 or higher nurseries in New York State we found that 1.9% of infants were < 34.5 (cold), 9.6% above 34.5 but < 35.5 (very cool), 48.0% above 35.5 but < 36.5 (cool), 37.9% above 36.5 but < 37.5 (about right or appropriately warm), and 2.6% above 37.5.
- The distribution of mean temperature by nursery ranged from 35.7 to 38.2, with a median of 36.3, a standard error of 0.36, and an interquartile range of 0.4.
- Twenty-five percent of these nurseries had a mean temperature below 36.1.

Disparities

Additionally, the developer report findings on disparities from their testing:

- Racial differences were seen in the New York State neonatal data analysis with black babies most likely to be cold, very cool, or cool and least likely to be about right or above normal (p<.001).
- Whites were least likely to be cool with non-Hispanic other and Hispanic infants at intermediate values.
- Race and ethnicity were also independent predictors of temperature in our New York City data.

Questions for the Committee:

○ *Is there a gap in care that warrants a national performance measure?*

○ *Should this measure be used to identify disparities in this area of healthcare?*

Preliminary rating for opportunity for improvement: ☒ **High** ☐ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 1: Importance to Measure and Report (including 1a, 1b, 1c)

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability

2a1. Reliability Specifications

2a1. Specifications requires the measure, as specified, to produce consistent (reliable) and credible (valid) results about the quality of care when implemented. The measure should be well defined and precisely specified so it can be implemented consistently within and across organizations and allows for comparability.

Data source(s): Administrative claims, Electronic Clinical Data : Electronic Health Record, Other, Paper Medical Records

Specifications:

- The levels of analysis are facility, health plan, integrated system, population (state, country, region).
- The developer states "***This measure does not have the form of numerator and denominator. It is a distribution. We ask for reporting of the distribution in terms of five categories across the distribution, in terms of key moments in the distribution, and as a graphical presentation of the distribution.*** " See example.
- The distribution of values as follows:

Category 1	Cold	≤ 34.5° C
Category 2	Very cool	>34.5 to ≤35. 5° C
Category 3	Cool	>35.5 to ≤36. 5° C
Category 4	About right	>36.5 to ≤37. 5° C
Category 5	Overly warm	>37. 5° C

The percent in each category is reported to 1 decimal places and is calculated as the number of temperatures meeting the criteria for that category divided by the denominator, that value multiplied by 100.

- The value to be reported is the first temperature upon arrival to the nursery. The numerator data elements include 1)Temperature to first decimal place; 2) Units of temperature (Celsius, Fahrenheit. Those measured in Fahrenheit should be converted to Celsius.); 3)Time that temperature was measured; 4)Time of arrival to the nursery (not time that admission was done); Optional: 5)Method of temperature measurement (axillary, rectal, skin, tympanic)
- The numerator specifications also state ‘‘Accountability entities may choose to use any of various components for their emphasis (alone or in combination), including percent "about right", mean or median temperatures, or value of the 10th or 25th percentiles, and the inter-percentile range."
- The developer also includes reporting of non-categorical description of the distribution including Descriptors of the Center of the Distribution (for sample size >5); Descriptors of Dispersion; Descriptors of the Warm End of the Distribution; and Descriptors of the Cool End of the Distribution.
- The developer acknowledges that "an accountability entity may prefer to summarize these categories as a single number. In such cases we recommend adding the number in Category 1, Category 2, and Category 3 and dividing that sum by the denominator and multiplying that proportion by 100 to calculate the percent. We term this percent to represent the number that are "Too Cool". "
- The denominator is "Live-born neonates with birthweight of less than 2,500 grams (as identified from either the medical record or by ICD-10-CM Principal or Other Diagnosis Codes in Table 1) admitted to a Level 2 or higher

nursery within 24 hours of birth. ICD-09-CM Codes are available in Table 1 for time frames that have not yet converted to ICD-10-CM. Children identified as having received Level 2 care either via medical record review and/or via revenue code 172, 173, or 174 shall be eligible for the denominator. "

- Exclusions for this measure include:
  - Neonates with anencephaly
  - neonates who receive only comfort care in the Level 2 or higher nursery,
  - neonates who die or are placed intentionally on a pre-existing hypothermia protocol prior to the 15 minute after arrival specification time.
- Stratification: the developer indicates that the measure "results must be stratified in three ways:
  - Main Stratum: Time between arrival at Level 2 or higher nursery is between 0 and 15 minutes.
  - Delayed stratum: Time between arrival at Level 2 or higher nursery is more than 15 minutes.
  - Other: Inadequate documentation to determine timing of temperature."
- Stratification: the developer also recommends stratifying by race/ethnicity; three birthweight categories; rural/ urban; poverty level and several other optional data elements
- Risk-adjustment "Stratification by risk category/subgroup". The developer states that risk adjustment by stratification: "For improvement and accountability purposes stratification by weight class and by race/ethnicity are specified."
- In **S16. type of score** the developer indicates "continuous, e.g., average" and indicates the interpretation is "Better quality = Score within a defined interval"
- A calculation algorithm is presented.

Questions for the Committee:

- *Is a distribution result appropriate for an accountability measure?*
- *Are the categories in the distribution appropriate?*
- *Is the single number result sum of categories 1-3 "Too Cool" more appropriate for an accountability measure?*
- *Are all the data elements clearly defined? Are all appropriate codes included?*
- *Is it clear what stratifications are required?*
- *Are the required stratifications appropriate and meaningful?*
- *Is the logic or calculation algorithm clear?*
- *There seem to be various approaches to the measure results (distribution, sum of categories 1-3, optional stratifications). Are the measure specifications precise and unambiguous such that it is likely this measure will be consistently implemented?*

2a2. Reliability Testing Testing attachment

2a2. Reliability testing demonstrates if the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise enough to distinguish differences in performance across providers.

SUMMARY OF TESTING

Reliability testing level ☐ **Measure score** ☒ **Data element** ☐ **Both**

Reliability testing performed with the data source and level of analysis indicated for this measure ☐ **Yes** ☐ **No**

Method(s) of reliability testing

- Literature review of guidelines and four studies to support "As a vital sign of longstanding, temperature is familiar and well recorded" and "route of measurement is not critical in this population premature infants and that commonly used methods to quantify temperature are sufficiently reliable. "
- Evidence published in a 2006 doctoral dissertation by R. Knobel (page 41) evaluating inter-rater reliability of nurse observation of neonatal temperature.
- A feasibility study designed to determine the ability and ease of collecting related data.

Results of reliability testing

- The Cohen's kappa for neonatal temperature in Knobel's work is 0.96. Kappa is intended to provide a quantitative measure of the magnitude of the agreement. A Kappa value of 0.81 – 0.99 indicates almost perfect agreement.
- A feasibility study "showed that date and time are self-evident and that there is mild but manageable variation in how time is reported."

Guidance from the Reliability Algorithm

(Box 1) Precise, unambiguous and complete specifications so that they can be consistently implemented – many options →low

Specific to testing only: empiric reliability testing (Box 2) →data elements testing (Box 8) →appropriate method; were ALL critical data elements tested? → high/moderate certainty that the data used are reliable → MODERATE (In the absence of reliability testing of the measure score, the highest rating possible is moderate)

Questions for the Committee:

- *Were ALL critical data elements tested?*
- *Is the test sample adequate to generalize for widespread implementation?*
- *Do the results demonstrate sufficient reliability so that differences in performance can be identified?*

Preliminary rating for reliability: ☐ **High** ☐ **Moderate** ☒ **Low** ☐ **Insufficient**

2b. Validity

2b1. Validity: Specifications

2b1. Validity Specifications. This section should determine if the measure specifications are consistent with the evidence.

Specifications consistent with evidence in 1a. ☐ **Yes** ☒ **Somewhat** ☐ **No**

Specification not completely consistent with evidence

- The temperature categories used in the distribution were determined by the developer rather than evidence.
- The "Too Cool" sum of categories 1-3 is consistent with the WHO guidelines temperature target range.

Question for the Committee:

- *Are the specifications consistent with the evidence?*

2b2. Validity testing

2b2. Validity Testing should demonstrate the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality.

SUMMARY OF TESTING

Validity testing level ☐ **Measure score** ☒ **Data element testing against a gold standard** ☐ **Both**

Method of validity testing of the measure score:

- ☒ **Face validity only**
- ☒ **Empirical validity testing of the measure score**

Validity testing method:

1. Data element testing of the ICD-9 codes for low birthweight.
2. Face validity of the experts and methods used to develop the measure.
3. Construct validity testing of the measure score for three hospitals in NY comparing temperature to mortality.

Validity testing results:

1. 99 infants out of 677 who were identified with ICD-9 specifications listed in Table 1, Section I, had birthweights of over 2,500. (14.6%) The developers note that "specifications of the various ICD-9 codes, such as 764.00, 764.10, and 765.10 that represent poor fetal growth without a specified weight need to have their eligibility for the measure confirmed with an actual birthweight."
2. The developers conclude "Temperature is a valid and meaningful outcome. The categorical and continuous descriptions of the distribution of temperatures are each valid."
3. The study of three hospitals found an association of hospital with lower temperatures and of lower temperatures with mortality were all significant using a generalized linear model approach at p<.01. " Using the proportion very cool or cold:

	Proportion Cold and Very Cool (categories 1 and 2)	Mortality
Hospital A	0.344	0.051
Hospital B	0.381	0.104
Hospital C	0.414	0.131

Questions for the Committee:

- *The construct validity analysis used the temperatures from categories 1 and 2 – this is another version of this measure. Testing should demonstrate validity of the measure as specified. Is the version of the measure tested the same as the measure submitted for possible endorsement?*
- *Is the test sample adequate to generalize for widespread implementation?*
- *Do the results demonstrate sufficient validity so that conclusions about quality can be made?*
- *Do you agree that the score from this measure as specified is an indicator of quality?*

2b3-2b7. Threats to Validity

2b3. Exclusions:

- The developer reports "Because the exclusions represent a distinct population with a higher risk of short term death, such analyses would be inappropriate. In our three hospital study, we excluded 31 of 746 (4.15%).

Questions for the Committee:

- *Are the exclusions consistent with the evidence?*
- *Are any patients or patient groups inappropriately excluded from the measure?*
- *Are the exclusions/exceptions of sufficient frequency and variation across providers to be needed (and outweigh the data collection burden)?*

2b4. Risk adjustment: **Risk-adjustment method** ☐ **None** ☐ **Statistical model** ☒ **Stratification**

Conceptual rationale for SDS factors included? ☒ **Yes** ☐ **No**

SDS factors included in risk model? ☒ **Yes** ☐ **No**

Risk adjustment summary

- "For improvement and accountability purposes stratification by weight class and by race/ethnicity are specified." However, this is not explicit in S12. Of the specifications.
- The developer provides stratified results for race/ethnicity and birthweight but states that "this is not relevant for our stratification. That is not its purpose. Its purpose is to report additional information. "

Questions for the Committee:

- *Is an appropriate risk-adjustment strategy included in the measure?*
- *Are the candidate and final variables included in the risk adjustment model adequately described for the measure to be implemented?*
- *Are all of the risk adjustment variables present at the start of care?*

2b5. Meaningful difference *(can statistically significant and clinically/practically meaningful differences in performance measure scores can be identified)*:

- The developer reports "The distribution of mean temperature by hospital ranged from 35.7 to 38.2, with a median of 36.3, a standard error of 0.36, and an interquartile range of 0.4. Twenty-five percent had a mean temperature below 36.1. We conclude that temperatures do vary across hospitals, with a range of ~7 standard errors. This is statistically meaningful variation."

Question for the Committee:

- *This data reports on hospital mean temperature – this is another version of the measure concept – but it is not included in the specifications. Is this a more meaningful measure result (mean hospital temperature) to compare hospitals?*
- *How are meaningful differences determined using the distribution results specified in S4 of the specifications?*
- *Does this measure identify meaningful differences about quality?*

2b6. Comparability of data sources/methods: Developer did not respond to this question.

- The developer indicated the following data sources:
  - Medical record abstraction (paper or EHR)
  - Administrative claims
  - Survey of hospital quality team
  - State repository of data

Question for the Committee:

- *Will use of a variety of data sources lead to comparable results that will allow fair comparisons among hospitals?*

2b7. Missing Data

- The developer states "Given the importance of temperature measurement in this population, there should be no missing temperatures or body weights. When calculating the distribution of temperatures, all missing temperatures should be considered as if the recorded temperature was 34 degrees. For the stratifications, the infants should be omitted and the number omitted reported."

Guidance from the algorithm

Specifications aligned with the evidence (multiple version/potential reliability issues (Box 1) → low

Threats to validity addressed (Box 2)– issues with risk adjustment, comparability, meaningful differences → insufficient

Empirical validity testing (Box 3)using the measure as specified – no (see construct validity) → face validity →moderate (in the absence of proper empiric validity testing the highest rating possible is moderate

Preliminary rating for validity: ☐ **High** ☐ **Moderate** ☐ **Low** ☒ **Insufficient**

Committee pre-evaluation comments

Criteria 2: Scientific Acceptability of Measure Properties (including all 2a, 2b, and 2d)

Criterion 3. Feasibility
Maintenance measures – no change in emphasis – implementation issues may be more prominent

3. Feasibility is the extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

- The developer describes several feasibility assessments for the data.
- The developer reports" The data required for the CAPQuaM perinatal measures are generally available in the existing data systems. We cannot comment on the readiness of systems to provide routine output into a database suitable for analysis and generation of these measures, but there are not fundamental barriers to such being accomplished. "
- "We have developed a stand-alone web-based data entry portal that supports this measure for all infants admitted to a Level 2 or higher nursery. We have also designed an interface that works with current EHR systems (Epic) at Mount Sinai to collect the necessary data as a part of normal workflow."

Questions for the Committee:

- *Are the required data elements routinely generated and used during care delivery?*
- *Are the required data elements available in electronic form, e.g., EHR or other electronic sources?*
- *Is the data collection strategy ready to be put into operational use?*

Preliminary rating for feasibility: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**
IF clear, unambiguous specifications are used

Committee pre-evaluation comments
Criteria 3: Feasibility

Criterion 4: Usability and Use
Maintenance measures – increased emphasis – much greater focus on measure use and usefulness, including both impact /improvement and unintended consequences

4. Usability and Use evaluate the extent to which audiences (e.g., consumers, purchasers, providers, policymakers) use or could use performance results for both accountability and performance improvement activities.

Current uses of the measure [from OPUS]
Publicly reported? ☐ **Yes** ☒ **No**

Current use in an accountability program? ☐ **Yes** ☒ **No**
OR
Planned use in an accountability program? ☐ **Yes** ☐ **No** unclear

Accountability program details none
The only use of this new measure is "The measure is in the process of being implemented for routine quality measurement at the Mount Sinai Medical Center. "

The developer provides an example of how the measure might be reported.

Improvement results NA

Unexpected findings (positive or negative) during implementation The developer reports "There has not been any evidence of unintended negative consequences to individuals or populations. There are no anticipated unintended consequences if measuring at the level of comparing states, geographic regions, payment models, or health plans. When comparing hospitals it may be important to incorporate strata that define context, such as rurality, poverty, and types of insurance."

Potential harms none specified

Feedback: none

Questions for the Committee*:*

- *Will the distribution results provide easy to understand comparisons among hospitals for public reporting?*
- *Will consumers, patients, payers, policy makers be able to use the distribution results?*
- *Will the distribution results provide results appropriate for payment incentives or other accountability purposes?*
- *Are the temperature distribution results the most useful and meaningful measure result for accountability purposes?*
- *Would another version of the measure such as mean temperature or "Too Cool" be more useful as an accountability measure?*
- *How can the performance results be used to further the goal of high-quality, efficient healthcare?*
- *Do the benefits of the measure outweigh any potential unintended consequences?*

Preliminary rating for usability and use: ☐ **High** ☐ **Moderate** ☒ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 4: Usability and Use

Criterion 5: Related and Competing Measures

Related or competing measures

none

Harmonization

NA

Pre-meeting public and member comments

•

NATIONAL QUALITY FORUM—Evidence (subcriterion 1a)

Measure Number (*if previously endorsed*)**:** Click here to enter NQF number
Measure Title: Thermal Condition of Low Birthweight Neonates Admitted to Level 2 or Higher Nurseries in the First 24 Hours of Life: A PQMP Measure
IF the measure is a component in a composite performance measure, provide the title of the Composite Measure here: Click here to enter composite measure #/ title

Date of Submission: 3/22/2016

Instructions

- *For composite performance measures:*
  - *A separate evidence form is required for each component measure unless several components were studied together.*
  - *If a component measure is submitted as an individual performance measure, attach the evidence form to the individual measure submission.*
- Respond to all questions as instructed with answers immediately following the question. All information needed to demonstrate meeting the evidence subcriterion (1a) must be in this form. An appendix of *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 10 pages (*incudes questions/instructions*; minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.

Note: The information provided in this form is intended to aid the Steering Committee and other stakeholders in understanding to what degree the evidence for this measure meets NQF's evaluation criteria.

1a. Evidence to Support the Measure Focus
The measure focus is evidence-based, demonstrated as follows:

- Health outcome: [3] a rationale supports the relationship of the health outcome to processes or structures of care. Applies to patient-reported outcomes (PRO), including health-related quality of life/functional status, symptom/symptom burden, experience with care, health-related behavior.
- Intermediate clinical outcome: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured intermediate clinical outcome leads to a desired health outcome.
- Process: [5] a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured process leads to a desired health outcome.
- Structure: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured structure leads to a desired health outcome.
- Efficiency: [6] evidence not required for the resource use component.

Notes

3. Generally, rare event outcomes do not provide adequate information for improvement or discrimination; however, serious reportable events that are compared to zero are appropriate outcomes for public reporting and quality improvement.
4. The preferred systems for grading the evidence are the U.S. Preventive Services Task Force (USPSTF) grading definitions and methods, or Grading of Recommendations, Assessment, Development and Evaluation (GRADE) guidelines.
5. Clinical care processes typically include multiple steps: assess → identify problem/potential problem → choose/plan intervention (with patient input) → provide intervention → evaluate impact on health status. If the measure focus is one step in such a multistep process, the step with the strongest evidence for the link to the desired outcome should be selected as the focus of measurement. Note: A measure focused only on collecting PROM data is not a PRO-PM.
6. Measures of efficiency combine the concepts of resource use and quality (see NQF's Measurement Framework: Evaluating Efficiency Across Episodes of Care; AQA Principles of Efficiency Measures).

1a.1.This is a measure of: (*should be consistent with type of measure entered in De.1*)
Outcome

☐ Health outcome: Click here to name the health outcome

☐Patient-reported outcome (PRO): Click here to name the PRO

PROs include HRQoL/functional status, symptom/symptom burden, experience with care, health-related behaviors

☒ Intermediate clinical outcome (*e.g., lab value*): Thermal condition/hypothermia

☐ Process: Click here to name the process

☐ Structure: Click here to name the structure

☐ Other: Click here to name what is being measured

_____

HEALTH OUTCOME/PRO PERFORMANCE MEASURE *If not a health outcome or PRO, skip to 1a.3*

1a.2. Briefly state or diagram the path between the health outcome (or PRO) and the healthcare structures, processes, interventions, or services that influence it.

This measure addresses a key gap in inpatient perinatal care. Evidence that thermal management (such as hot water bottles and incubators) improves survival of newborn and premature infants exists from as early as the late 19th century (1-7). Modern studies have confirmed and extended these findings and this has been considered settled science since Silverman's classic paper in Pediatrics in 1958. The capacity of interventions to improve the thermal condition of neonates was also demonstrated by Silverman and more nuanced studies have confirmed the opportunity for management to improve outcomes. (8-10). Laptook et al not only reaffirmed the association of temperature loss with poor outcomes in 5,277 infants, 401-1,499 grams, born at any of 15 academic medical centers participating in the National Institute of Child Health and Development (NICHD) Neonatal Research Network, but also the existence of a serious care gap even in these elite nurseries, with nearly 2 in 5 (38.3%) 28 week gestation infants being hypothermic (less than 36 degrees Celsius); younger babies had worse thermal outcomes. Nearly one in ten (9.3% of the 28 week infants had temperatures below 35 Celsius (11). A formal item selection process looking at potential measures for infants under 1,500 grams identified neonatal temperature as an independent contributor to a composite quality of care measure (12).

Drs Kleinman, Howell and others at Mount Sinai collected chart review data from three diverse hospitals in New York City. All three hospitals had a range of birthweights and a range of temperatures, both when we considered the actual measured temperature and when we adjusted those that were not taken rectally to create a "corrected" core temperature. See Figures 1 and 2 below. Subsequent reading in the literature (including the WHO guideline) have convinced us that accepted best practice is to accept the measured temperature regardless of what route was used to measure the temperature.

Figure 1

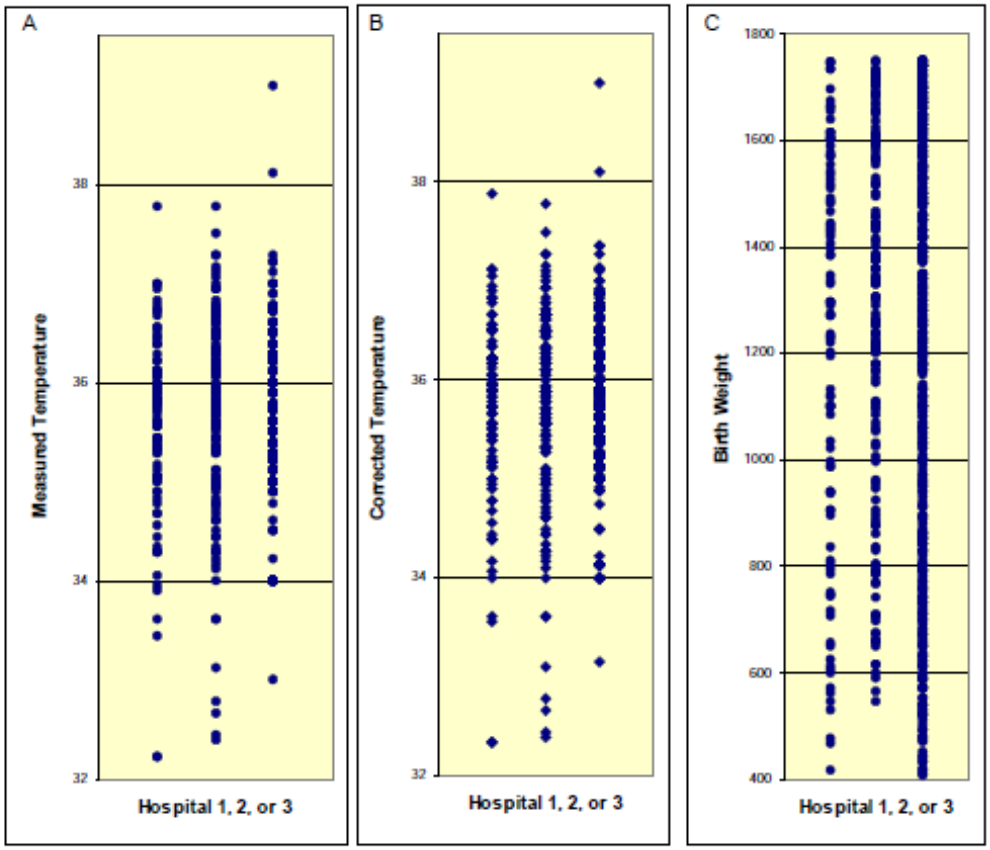


Figure 1 (A-C). Scatter plots of measured temperature (A), temperature adjusted for method of measurement (B), and birthweight in grams (C) for each of the three study hospitals. N = 100, 158, and 487 for Hospitals 1, 2, and 3 respectively.

Figure 2

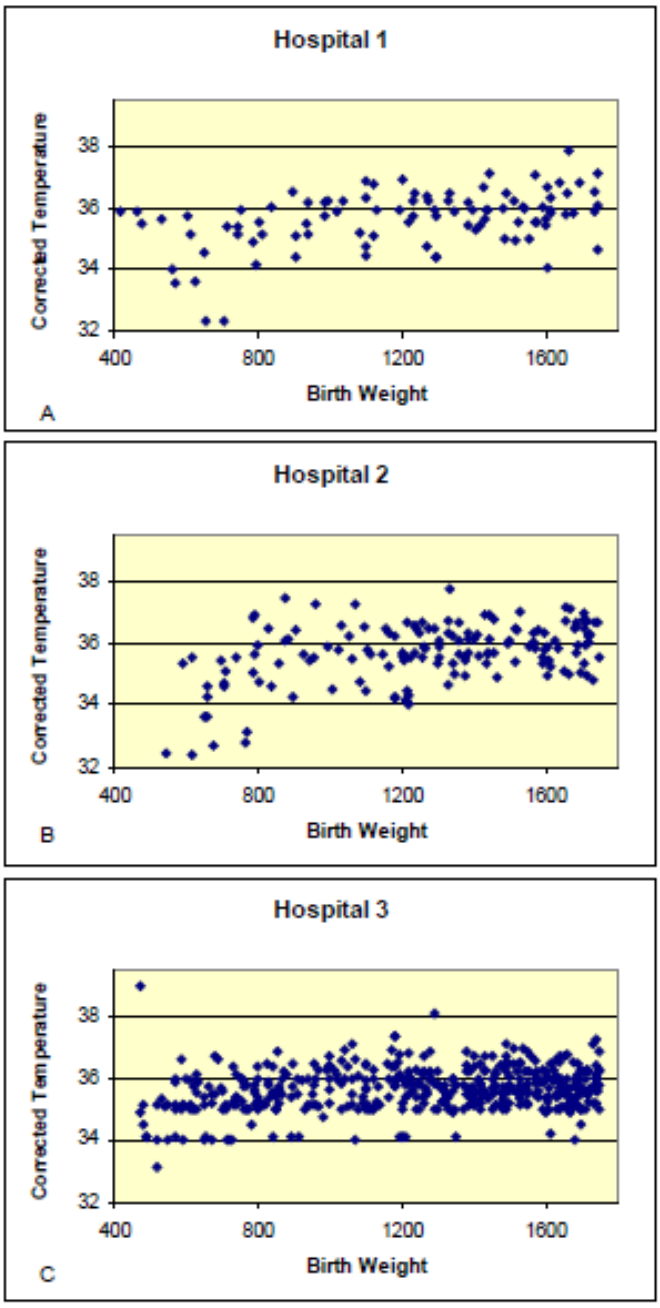


Figure 2 (A-C). Scatter plot of temperature by birthweight in grams for each of the three study hospitals. Valid measurements on both birthweight and temperature were available for 99 infants in Hospital 1, 147 infants in Hospital 2, and 459 infants in Hospital 3.

Temperature predicted in-hospital mortality after controlling for covariates, whether we dichotomized at the 35.5° threshold that our local physicians proposed or considered each degree of temperature as a continuous variable. Crossing the threshold into hypothermia more than doubled the odds of death, controlling for other variables in the model. The relationship between temperature and survival is monotonic: an increase of each 1° Celsius up to 37° reduced odds of death by more than 35% in the model using a continuous variable (22% for 1° Fahrenheit). Defining hypothermia as admission temperature below 36.0 would estimate an increase in the risk of mortality by 84%, p=0.19.

Risk ratio (RR) is a more informative way to express the results than an odds ratio especially when the underlying risk is large, as in this study (12). Regression risk analysis estimates the adjusted risk ratio (ARR) and adjusted risk difference: hypothermia (35.5C) results in an ARR of 1.48 (95% confidence interval 1.03—2.30), indicating a 48% increase in risk, from a baseline risk of 8.9% among those who were euthermic (about right) to an exposed risk of 13.1% among those who were

hypothermic, controlling for the covariates in the sample. Considering temperature as a continuous variable reveals that increasing the temperature from 34.0 to 35.0 increases the relative chance of survival by 24%, from 35.0 to 36.0 by 26%, and from 36.0 to 37.0 by 27%, resulting in absolute risk reductions of 2.8%, 2.4%, and 2.0% respectively. A core body temperature increase from 34.0 to 37.0 is associated with a relative decrease in mortality of 98% and an absolute decrease in mortality of 7.2%, controlling for other factors in the model. The decrease from 36.0 to 35.5 is associated with a 12% increase in the adjusted mortality risk from 9.4% to 10.5%. The impact of 1 degree Celsius warmer on survival is approximately equivalent to the infant having an additional 100 grams of birthweight.

Our work confirmed findings in the literature that insurance status and race (13) are associated with outcomes. Anecdotal reports from among our participating hospitals confirm reports in the literature that attention to thermal management can improve temperature outcomes (14). Our team published a detailed review of the literature (15).

References

1. Garrison FH. *History of Pediatrics* in Abt IA. Pediatrics. Volume I. WB Saunders. Philadelphia. 1923..
2. Holt LE. *The Care of Premature and Delicate Infants* in Holt LE. The Diseases of Infancy and Childhood, Second Edition Revised and Enlarged. D. Appleton. NY 1902
3. Baker, JP. The Incubator and the Medical Discovery of the Premature Infant. Journal of Perinatology 5:321-328. 2000.
4. Pierce RV. The People's Medical Adviser in Plain English. World's Dispensary Printing Office. Buffalo, NY 1875
5. Currier, A. *Diseases of the Newborn*. In Sajous CE. (editor). Annual of the Universal Medical Sciences. FA Davis, Philadelphia. 1891. Volume II, Section L.
6. Fischer L/ Diseases of Infancy and Childhood. FA Davis and Co. 1915.
7. Holt Jr. EL, McIntosh R. Holt's Diseases of Infants and Children, 11th Edition. Appleton. New York. 1940.
8. Silverman WA, Fertig JW, Berger AP. The influence of the thermal environment upon the survival of newly born premature babies. Pediatrics. 22:876-886. 1958.
9. Sinclair JC. Servo-control for maintaining abdominal skin temperature at 36C in low birth weight infants. *Cochrane Database of Systematic Reviews* 2007, Issue 4, Art. No. CD001074.
10. Watkinson M. Temperature Control of Premature Infants in the Delivery Room. Clinics in Perinatology 33:43-53.2006
11. Laptook AR, Salhab W, Bhaskar B, et al. Admission Temperature of Low Birth Weight Infants.; Predictors and Associated Morbidities. Pediatrics. 119:e643-649. 2007.
12. Howell EA, Stone J, Inamdar S, Matseoane S, Kleinman LC, Chassin MR. Approaching Guideline Recommended Care for Maternal-Infant Health: Clinical failures to use recommended antenatal corticosteroids, Maternal Child Health Journal, 2009 Jun 4. [Epub ahead of print]
13. Howell EA, Holzman I, Kleinman LC, Wang J, Chassin MR. Surfactant use for premature infants with respiratory distress syndrome in three New York city hospitals: discordance of practice from a community clinician consensus standard. Journal of Perinatology; 30:590-595. 2010
14. Chatterjee S, Price B. Regression Analysis By Example, Fourth Edition. John Wiley & Sons. New York. 2006.
15. Fastman-Rabin, Howell, Holzman & Kleinman. Current perspectives on temperature management and hypothermia in low birth weight infants. Newborn & Infant Nursing Reviews. June 2014. Vol. 14;2. 50-55.

1a.2.1. State the rationale supporting the relationship between the health outcome (or PRO) to at least one healthcare structure, process, intervention, or service (*i.e., influence on outcome/PRO*).

Temperature of low birthweight neonates is variable, can be managed at the level of the individual patient as well as at the level of the unit providing care, and is highly consequential in terms of critical outcomes such as survival and intraventricular hemorrhage. At a population level, the lower the temperature the larger the consequences.

Successful interventions that can reduce hypothermia include the use of radiant warmers, quality improvement collaboratives (1), raising temperature in the delivery and operating rooms (2-4), and the use of immediate wrapping of the newborn in any of several approaches to wrapping (5-20). Favorable thermal outcomes are amenable to management and are associated with specific processes of care. The scatterplots shown above from our work in NYC

hospitals demonstrate that even very small babies can be maintained at normal temperatures. Foundational work done at the turn of the 19th to the 20th century demonstrated that meticulous thermal management even with incubators or hot water bottles can improve outcomes.

A Cochrane review has confirmed the efficacy of plastic bags in addition to radiant warming in improving the NICU admission temperature of premature babies. (21)

References

1. Kaplan, HC., Lannon, C., Walsh, MC., Donovan, EF. Ohio statewide quality-improvement collaborative to reduce late-onset sepsis in preterm infants. Pediatrics. March 2011. 127 (3). 427-435.
2. World Health Organization (WHO). Thermal control of the newborn: a practical guide. Maternal Health and Safe Motherhood Programme (WHO/FHE/MSM/93.2). WHO: Geneva, 1996.
3. American Society of PeriAnesthesia Nurses. Clinical guideline for the prevention of unplanned perioperative hypothermia [online]. 2000 Oct [cited 2008 Jan 29]. Available from Internet: http://www.aspan.org/PDFfiles/HYPOTHERMIA_GUIDELINE10-02.pdf.
4. AORN Recommended Practices Committee. Recommended practices for the prevention of un¬planned perioperative hypothermia. AORN J 2007 May;85 (5):972-4, 976-84, 986-8.
5. Russo, A., McCready, M., Torres, L., et al. Reducing hypothermia in preterm infants following delivery. Pediatrics. April 2014. 133(4). 1055-1062.
6. Vohra S, Frent G, Campbell V, Abbott M, Whyte R. Effect of polyethylene occlusive skin wrapping on heat loss in very low birth weight infants at delivery: a randomized trial. J Pediatr. 1999;134(5):547–551.
7. Vohra S, Roberts RS, Zhang B, Janes M, Schmidt B. Heat Loss Prevention (HeLP) in the delivery room: a randomized controlled trial of polyethylene occlusive skin wrapping in very preterm infants. J Pediatr. 2004;145(6):750–753.
8. Knobel RB, Wimmer JE Jr, Holbert D. Heat loss prevention for preterm infants in the delivery room. J Perinatol. 2005;25(5):304–308.
9. Lee HC, Ho QT, Rhine WD. A quality improvement project to improve admission temperatures in very low birth weight infants. J Perinatol. 2008;28(11):754–758.
10. Kent AL, Williams J. Increasing ambient operating theatre temperature and wrapping in polyethylene improves admission temperature in premature infants. J Paediatr Child Health. 2008;44(6):325–331.
11. Ibrahim CP, Yoxall CW. Use of plastic bags to prevent hypothermia at birth in preterm infants—do they work at lower gestations? Acta Paediatr. 2009;98(2):256–260.
12. McCall EM, Alderdice F, Halliday HL, Jenkins JG, Vohra S. Interventions to prevent hypothermia at birth in preterm and/or low birthweight infants. Cochrane Database Syst Rev. 2010;3(3):CD004210
13. Singh A, Duckett J, Newton T, Watkinson M. Improving neonatal unit admission temperatures in preterm babies: exothermic mattresses, polythene bags or a traditional approach? J Perinatol. 2010;30(1):45–49.
14. Trevisanuto D, Doglioni N, Cavallin F, Parotto M, Micaglio M, Zanardo V. Heat loss prevention in very preterm infants in delivery rooms: a prospective, randomized, controlled trial of polyethylene caps. J Pediatr. 2010;156 (6):914–917, e1.
15. te Pas AB, Lopriore E, Dito I, Morley CJ, Walther FJ. Humidified and heated air during stabilization at birth improves temperature in preterm infants. Pediatrics. 2010;125(6). Available at: www.pediatrics.org/cgi/content/full/125/6/e1427.
16. Simon P, Dannaway D, Bright B, et al. Thermal defense of extremely low gestational age newborns during resuscitation: exothermic mattresses vs polyethylene wrap. J Perinatol. 2011;31(1):33–37.
17. Chawla S, Amaram A, Gopal SP, Natarajan G. Safety and efficacy of Trans-Warmer mattress for preterm neonates: results of a randomized controlled trial. J Perinatol. 2011;31(12):780–784

18. McCarthy LK, O'Donnell CPF. Warming preterm infants in the delivery room: polyethylene bags, exothermic mattresses or both? Acta Paediatr. 2011;100(12):1534– 1537.
19. Mathew B, Lakshminrusimha S, Sengupta S, Carrion V. Randomized controlled trial of vinyl bags versus thermal mattress to prevent hypothermia in extremely lowgestational-age infants. Am J Perinatol. 2013;30(4):317–322.
20. Billimoria Z, Chawla S, Bajaj M, Natarajan G. Improving admission temperature in extremely low birth weight infants: a hospitalbased multi-intervention quality improvement project. J Perinat Med. 2013;41(4):455–460.
21. McCall EM, Alderdice F, Halliday HL, Jenkins JG, Vohra S. Interventions to prevent hypothermia at birth in preterm and/or low birth weight infants. Cochrane Database of Systematic Reviews 2010, Issue 3. Art. No.: CD004210

Note: For health outcome/PRO performance measures, no further information is required; however, you may provide evidence for any of the structures, processes, interventions, or service identified above.

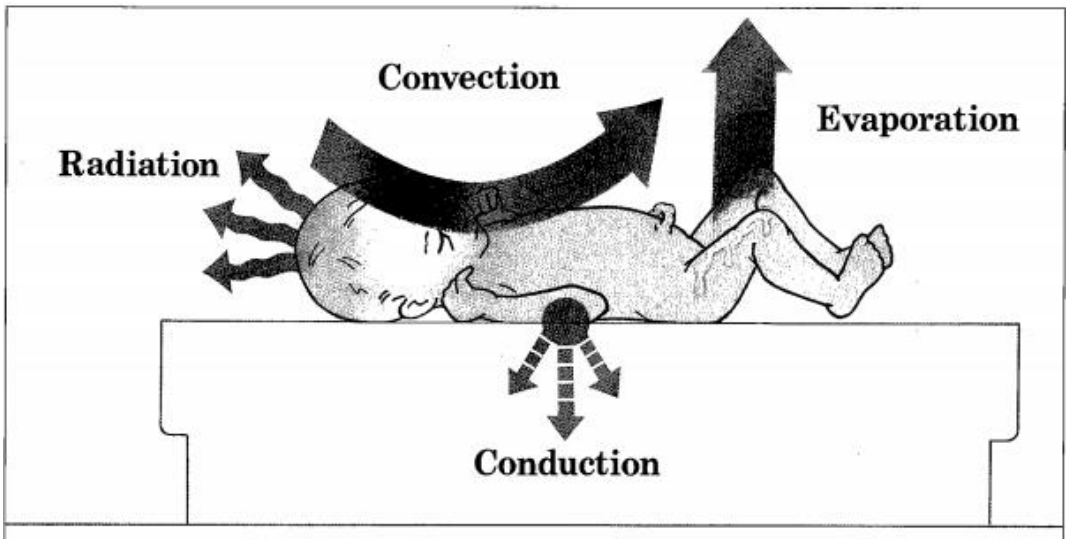
INTERMEDIATE OUTCOME, PROCESS, OR STRUCTURE PERFORMANCE MEASURE

1a.3. Briefly state or diagram the path between structure, process, intermediate outcome, and health outcomes. Include all the steps between the measure focus and the health outcome.

Multiple studies over the course of more than a century have shown the association between hypothermia and mortality. Our work and others have also demonstrated that intracranial hemorrhage is more common among infants who become cool than among those who remain euthermic (about right).

The temperature inside the mother's womb is 38 degrees Celsius (100.4 F). Leaving the warmth of the womb at birth, the wet newborn finds itself in much colder environment and immediately starts losing heat.

The newborn baby loses heat in four different ways. (1) evaporation of amniotic fluid from the baby's body, (2) conduction if they baby is placed naked on a cold surface, (3) convection, if the naked newborn is exposed to cooler surrounding air, and (4) radiation, from the baby to cooler objects in the vicinity. Heat loss increases with air movement. If heat loss is not prevented and is allowed to continue, the baby will develop hypothermia (e.g. body temperature below normal). The diagram below is from the World Health Organization (WHO) guideline.

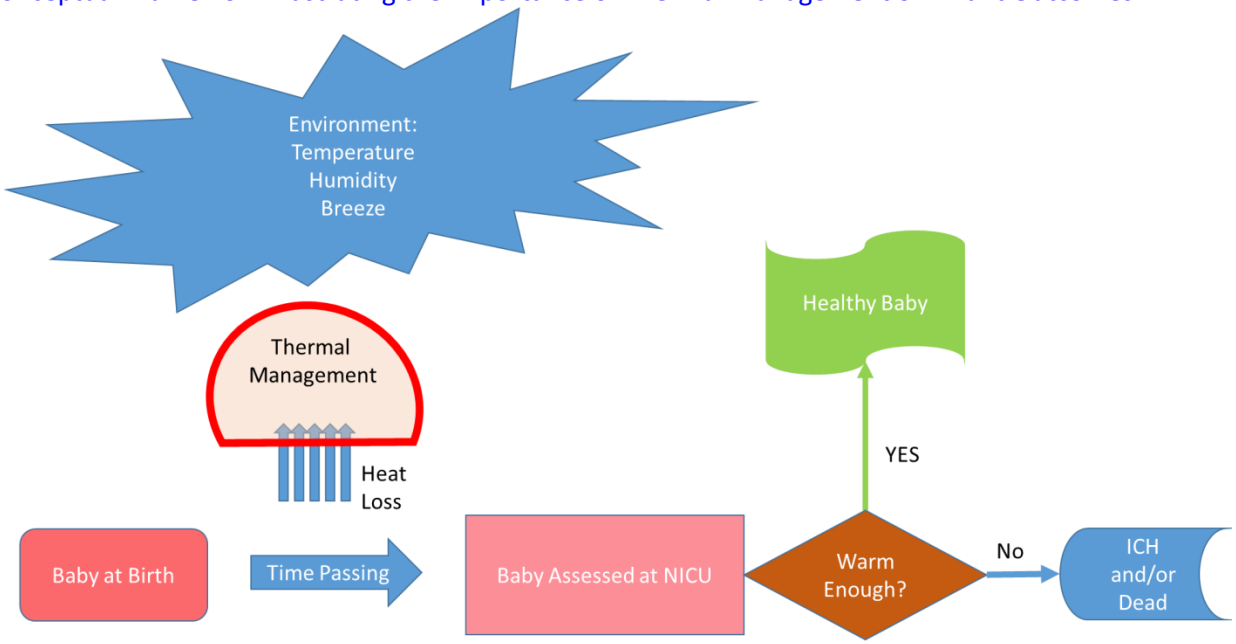


Four ways a newborn may lose heat to the environment

There is ample evidence that hypothermia is harmful. Prolonged hypothermia is linked to impaired growth and may make the new born more vulnerable to infections. Moreover, hypothermia, even if moderate, is associated with an increased risk of death in low birth weight newborns. Sick or low birth weight babies admitted to neonatal units with hypothermia are more likely to die than those admitted with normal temperatures. A hypothermic baby, especially if it is small or sick, is at increased risk of develping health problems and of dying.

Impact of Temperature on Mortality (Kleinman, Howell et al, not yet published)	
Change Temperature (Celsius)	Net Benefit to Mortality*
From 34 to 35	2.8%
From 35 to 36	2.4%
From 36 to 37	2.0%
From 34 to 36	5.2%
From 35 to 37	4.4%
From 34 to 37	7.2%
*All mortality differences are derived from an adjustment model that includes clincial and demographic factors including race/ethnicity and 5 minute Apgar score.	

Conceptual Framework Illustrating the Importance of Thermal Management on Infant Outcomes



Illustrating: Hypothermia can drive newborns down the pathway towards death or intracranial hemorrhage (ICH). Meticulous thermal management can keep neonates warm, increasing the likelihood of a healthy baby.

1a.3.1. What is the source of the systematic review of the body of evidence that supports the performance measure?

☒ Clinical Practice Guideline recommendation – ***complete sections 1a.4, and 1a.7***

☐ US Preventive Services Task Force Recommendation – ***complete sections 1a.5 and 1a.7***

☒ Other systematic review and grading of the body of evidence (*e.g., Cochrane Collaboration, AHRQ Evidence Practice Center*) – ***complete sections 1a.6 and 1a.7***

☐ Other – ***complete section 1a.8***

Please complete the sections indicated above for the source of evidence. You may skip the sections that do not apply.

_____

1a.4. CLINICAL PRACTICE GUIDELINE RECOMMENDATION

1a.4.1. Guideline citation (*including date*) and **URL for guideline** (*if available online*):

Multiple guidelines are available. Exemplar guidelines include:

World Health Organization
http://www.newbornwhocc.org/2014_pdf/Thermal%20management%202014.pdf

European Consensus Guidelines on the Management of Neonatal Respiratory Distress Syndrome in Preterm Infants – 2013 Update
http://bayanbox.ir/view/3068761397631809685/RDS-consensus.pdf

WAHT-NEO-048 This guideline has been printed from the Worcestershire Acute Hospitals NHS Trust intranet on 11/08/2015,15:29 It is the responsibility of every individual to check that this is the latest version/copy of this document.

https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=2&cad=rja&uact=8&ved=0ahUKEwjB6MinlfXKAhXEMj4KHfg-DpsQFggnMAE&url=http%3A%2F%2Fwww.worcsacute.nhs.uk%2FEasysiteWeb%2Fgetresource.axd%3FAssetID%3D23888%26type%3Dfull%26servicetype%3DAttachment&usg=AFQjCNEKw34TyBFuKl6fwemRMny7coxenA&sig2=f12lk48Fs_PF5WTlakY-oQ

1a.4.2. Identify guideline recommendation number and/or page number and **quote verbatim, the specific guideline recommendation**.

This is fully established science. There is no controversy regarding this linkage.

WHO:
A newborn baby is homeothermic, but his ability to maintain his body temperature can be easily overwhelmed by environmental temperatures. Thermal protection of the newborn is a set of continuing measures, which starts at birth, to ensure that he maintains a body temperature of 36.5°C to 37.5°C.

The WHO guideline states, hypothermia "is harmful to newborn babies, increasing the risk of illness and death." (p.1)

European Consensus Guidelines:
(5) Plastic bags or occlusive wrapping under radiant warmers should be used during stabilization in the delivery suite for babies <28 weeks gestation to reduce the risk of hypothermia.
(6) Babies stabilized under a radiant warmer should be servo-controlled within 10 minutes to avoid overheating.

WAHT NEO-048:
This Trust accepts that hypothermia in neonates is defined as temperature below 36.5ºC.

1a.4.3. Grade assigned to the quoted recommendation <u>with definition</u> of the grade:

WHO does not assign grades.

1a.4.4. Provide all other grades and associated definitions for recommendations in the grading system. (*Note: If separate grades for the strength of the evidence, report them in section 1a.7.*)

WHO does not assign grades.

1a.4.5. Citation and URL for methodology for grading recommendations (*if different from 1a.4.1*)**:**

1a.4.6. If guideline is evidence-based (rather than expert opinion), are the details of the quantity, quality, and consistency of the body of evidence available (e.g., evidence tables)?

☒ Yes → ***complete section 1a.7***

☐ No → ***report on another systematic review of the evidence in sections 1a.6 and 1a.7; if another review does not exist, provide what is known from the guideline review of evidence in 1a.7***

1a.5. UNITED STATES PREVENTIVE SERVICES TASK FORCE RECOMMENDATION

1a.5.1. Recommendation citation (*including date*) and **URL for recommendation** (*if available online*):

1a.5.2. Identify recommendation number and/or page number and **quote verbatim, the specific recommendation**.

1a.5.3. Grade assigned to the quoted recommendation with definition of the grade:

1a.5.4. Provide all other grades and associated definitions for recommendations in the grading system. (*Note: the grading system for the evidence should be reported in section 1a.7.*)

1a.5.5. Citation and URL for methodology for grading recommendations (*if different from 1a.5.1*)**:**

Complete section 1a.7

1a.6. OTHER SYSTEMATIC REVIEW OF THE BODY OF EVIDENCE

1a.6.1. Citation (*including date*) and **URL** (*if available online*):

1a.6.2. Citation and URL for methodology for evidence review and grading (*if different from 1a.6.1*)**:**

Complete section 1a.7

1a.7. FINDINGS FROM SYSTEMATIC REVIEW OF BODY OF THE EVIDENCE SUPPORTING THE MEASURE

If more than one systematic review of the evidence is identified above, you may choose to summarize the one (or more) for which the best information is available to provide a summary of the quantity, quality, and consistency of the body of evidence. Be sure to identify which review is the basis of the responses in this section and if more than one, provide a separate response for each review.

1a.7.1. What was the specific structure, treatment, intervention, service, or intermediate outcome addressed in the evidence review?

The thermal chain of management to keep infants warm. This guideline was field tested in 8 countries and the revised guideline incorporates lessons learned from its use. The underlying science is settled and not controversial.

1a.7.2. Grade assigned for the quality of the quoted evidence with definition of the grade:

Cochrane Review 1 (Sinclair): Randomized control trial evidence indicates that Servo-control of abdominal skin temperature reduces neonatal death rate in LBQ infants. If thermal neutrality is the goal incubator heating should be adjusted to maintain abdominal skin temperature at approximately 36.5 degrees Celsius.

Cochrane Review 2 (McCall): Reported on six studies, a totally of 304 infants randomized in the study. The authors conclude that there are a variety of methods that can keep preterm infants warmer on admission to neonatal units with less hypothermia. The number needed to treat was 2 with a range of 2-4.

- Sinclair JC. Servo-control for maintaining abdominal skin temperature at 36C in low birth weight infants. *Cochrane Database of Systematic Reviews* 2007, Issue 4, Art. No. CD001074.

- McCall EM, Alderdice F, Halliday HL, Jenkins JG, Vohra S. Interventions to prevent hypothermia at birth in preterm and/or low birthweight infants. Cochrane Database Syst Rev. 2010;3(3):CD004210

1a.7.3. Provide all other grades and associated definitions for strength of the evidence in the grading system.

1a.7.4. What is the time period covered by the body of evidence? (*provide the date range, e.g., 1990-2010*). Date range: Sinclair used data from 1964; McCall 1968-2005

QUANTITY AND QUALITY OF BODY OF EVIDENCE

1a.7.5. How many and what type of study designs are included in the body of evidence? (*e.g., 3 randomized controlled trials and 1 observational study*)

6 studies were identified that met the criteria of randomized study.

1a.7.6. What is the overall quality of evidence across studies in the body of evidence? (*discuss the certainty or confidence in the estimates of effect particularly in relation to study factors such as design flaws, imprecision due to small numbers, indirectness of studies to the measure focus or target population*)

This is established science with studies producing consistent results that meticulous thermal management decreases hypothermia and resultant bad outcomes.

ESTIMATES OF BENEFIT AND CONSISTENCY ACROSS STUDIES IN BODY OF EVIDENCE

1a.7.7. What are the estimates of benefit—magnitude and direction of effect on outcome(s) across studies in the body of evidence? (*e.g., ranges of percentages or odds ratios for improvement/ decline across studies, results of meta-analysis, and statistical significance*)

Large benefit. See data chart in 1A.3 quantify impact. Also see 1b.4 for risk ratios.

1a.7.8. What harms were studied and how do they affect the net benefit (benefits over harms)?

Depending on method chosen, there are risks of burns or overwarming. These deleterious side effects do not approach the magnitude of benefit from the systematic and meticulous thermal management of infants. These side effects may help to guide institutions to choose one over another of the effective methods for maintaining infant temperatures. This measure is agnostic to the process of thermal management and focuses instead on the outcome. The measure will identify those who are overheated.

UPDATE TO THE SYSTEMATIC REVIEW(S) OF THE BODY OF EVIDENCE

1a.7.9. If new studies have been conducted since the systematic review of the body of evidence, provide for each new study: 1) citation, 2) description, 3) results, 4) impact on conclusions of systematic review.

n/a

1a.8 OTHER SOURCE OF EVIDENCE

If source of evidence is NOT from a clinical practice guideline, USPSTF, or systematic review, please describe the evidence on which you are basing the performance measure.

1a.8.1 What process was used to identify the evidence?

1a.8.2. Provide the citation and summary for each piece of evidence.

1. Evidence, Performance Gap, Priority – Importance to Measure and Report

Extent to which the specific measure focus is evidence-based, important to making significant gains in healthcare quality, and improving health outcomes for a specific high-priority (high-impact) aspect of healthcare where there is variation in or overall less-than-optimal performance. ***Measures must be judged to meet all subcriteria to pass this criterion and be evaluated against the remaining criteria.***

1a. Evidence to Support the Measure Focus – See attached Evidence Submission Form

Thermal_Condition_evidence_attachment_Submission_03_22_16.docx

1b. Performance Gap

Demonstration of quality problems and opportunity for improvement, i.e., data demonstrating:

- considerable variation, or overall less-than-optimal performance, in the quality of care across providers; and/or
- disparities in care across population groups.

1b.1. **Briefly explain the rationale for this measure** *(e.g., the benefits or improvements in quality envisioned by use of this measure)*
Inpatient perinatal care was assigned to CAPQuaM as a PQMP priority by the Agency for Healthcare Research and Quality with the active consultation of the Centers for Medicare & Medicaid Services. After initial assignment, conversations between CAPQuaM, AHRQ, and CMS resulted in a decision for CAPQuaM to undertake the development of measures related to the temperature of low birthweight neonates. We developed this measure in close consultation with our Consortium partners at the New York State Department of Health, including the Office of Health Insurance Program/New York State Medicaid and the Division of Family Health.

Hypothermia in neonates has been identified as associated with poor outcomes since the 19th century and is a firmly established construct. We note that there is evidence that management can enhance thermal outcomes. In work that preceded the CAPQuaM work, our team demonstrated that the temperature of low birthweight neonates is variable, and is highly consequential in terms of critical outcomes like survival and intraventricular hemorrhage. Institutional anecdotal evidence supports literature observations that thermal management can be managed and improved at the unit level with improved outcomes.

1b.2. **Provide performance scores on the measure as specified (current and over time) at the specified level of analysis.** *(This is required for endorsement maintenance. Include mean, std dev, min, max, interquartile range, scores by decile. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included). This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*
In our study of 7,553 neonates admitted to Level 2 or higher nurseries in New York State we found that 1.9% of infants were < 34.5 (cold), 9.6% above 34.5 but < 35.5 (very cool), 48.0% above 35.5 but < 36.5 (cool), 37.9% above 36.5 but < 37.5 (about right or appropriately warm), and 2.6% above 37.5. The distribution of mean temperature by nursery ranged from 35.7 to 38.2, with a median of 36.3, a standard error of 0.36, and an interquartile range of 0.4. Twenty-five percent of these nurseries had a mean temperature below 36.1. Key findings from these analyses were: temperature was variable within weight categories; blacks were disproportionately cool compared with Hispanic or non-Hispanic others, who were disproportionately cool compared with non-Hispanic whites; and deaths were disproportionate among those who were cool, in a graded fashion. Only 36% of Medicaid infants were about right, compared to 40% of commercially insured infants. We also found systematic differences in the timing of when the temperatures were taken.

1b.3. **If no or limited performance data on the measure as specified is reported in 1b2, then provide a summary of data from the literature that indicates opportunity for improvement or overall less than optimal performance on the specific focus of measurement.**

n/a

1b.4. **Provide disparities data from the measure as specified (current and over time) by population group, e.g., by race/ethnicity, gender, age, insurance status, socioeconomic status, and/or disability.** *(This is required for endorsement maintenance. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities include.) This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*
Our feasibility assessment confirmed that racial and ethnicity data are almost universally available and that method of assignment of race and ethnicity to the baby varied. Assignment could be based on maternal self-report or assigned by the hospital, most typically as the mother's race and ethnicity. National improvement is needed in the methods used to assign race and ethnicity to newborns in the hospital. For the purposes of this measure, we are resigned at this time to using the existing data as recorded in the infants' medical records.

Racial differences were seen in our New York State neonatal data analysis with black babies most likely to be cold, very cool, or cool and least likely to be about right or above normal (p<.001). Whites were least likely to be cool with non-Hispanic other and Hispanic infants at intermediate values. Race and ethnicity were also independent predictors of temperature in our New York City data.

Our 3 hospital study included 715 infants of which 35.9% had a temperature below 35.5 degrees. The mean birth weight was 1242 grams (sd=356) and 9.7% of these infants died in the hospital. 25.5% were white, 34.3% black, and 31.8% Hispanic. 45.9% had public insurance, the others were privately insured.

In a fully risk adjusted logistic regression model that included potential confounders including 5 minute Apgar score (categorical), timing of delivery, presence of hyaline membrane disease, insurance status, race/ethnicity, birth weight and temperature, key findings included:

a. Each increase of 1 degree Celsius was associated with a 36% risk of mortality (ARR 0.64, 95% CI was 0.45-0.91, P<0.01);

b. Black race was associated with a more than tripling of the risk of mortality (ARR 3.32 (1.21-9.06, P<.0001)

c. Hispanic ethnicity was associated with a non-significant increase in mortality (ARR 1.11, 0.36-3.47, p=0.85)

d. Private insurance was associated with lower mortality (ARR 0.15, 0.07-0.36, P<.001)

1b.5. **If no or limited data on disparities from the measure as specified is reported in 1b4, then provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement. Include citations.**
n/a

1c. High Priority (previously referred to as High Impact)

The measure addresses:

- a specific national health goal/priority identified by DHHS or the National Priorities Partnership convened by NQF; OR
- a demonstrated high-priority (high-impact) aspect of healthcare (e.g., affects large numbers of patients and/or has a substantial impact for a smaller population; leading cause of morbidity/mortality; high resource use (current and/or future); severity of illness; and severity of patient/societal consequences of poor quality).

1c.1. **Demonstrated high priority aspect of healthcare**
Affects large numbers, A leading cause of morbidity/mortality, Frequently performed procedure, High resource use, Patient/societal consequences of poor quality, Severity of illness
1c.2. **If Other:**

1c.3. **Provide epidemiologic or resource use data that demonstrates the measure addresses a high priority aspect of healthcare. List citations in 1c.4.**
This measure addresses a key gap in inpatient perinatal care. Evidence that thermal management (such as hot water bottles and incubators) improves survival of newborn and premature infants exists from as early as the late 19th century (1-7). Modern studies have confirmed and extended these findings, including potential methods to maintain temperature for infants in the delivery room (8-10). Laptook et al confirmed the association of temperature loss with poor outcomes in 5,277 infants, 401-1,499 grams, born at any of 15 academic medical centers participating in the National Institute of Child Health and Development (NICHD) Neonatal Research Network (11). A formal item selection process looking at potential measures for infants under 1,500 grams identified

neonatal temperature as an independent contributor to a composite quality of care measure (12).

In New York State, about half of low birthweight babies are insured by Medicaid. Hypothermia is not only associated with neonatal mortality, but there is evidence that Intraventricular Hemorrhage (IVH) can also be a consequence of hypothermia (13). IVH is a significant cause of disability, developmental delay, and, when serious, is a common cause for LBW infants to develop into children with special health care needs. This has broad impact on Medicaid, Medicaid expenses, and early intervention services, including EPSDT services. Hypothermia, through death and disability, may have a long tail that impacts families and programs associated with Medicaid. Furthermore, the Medicaid population is disproportionately black and in our testing data, black infants were disproportionately hypothermic.

In addition to literature conducted in a variety of settings including the NICHD neonatal research network and the Vermont Oxford Network that document this problem, we have found performance concerns in New York City and New York State. Our chart review data from three diverse hospitals in New York City showed variation in temperatures recorded across the weight spectrum within and between hospitals. These differences were meaningful with cooler babies more likely to die. The importance of evaluating the spectrum of temperature is evident from our analyses with temperature as a continuous variable. These analyses reveal that each increase in degree of temperature increases the relative chance of survival significantly.

This history, these data, and the absence of currently recommended measures that adequately address this issue all motivate the work of the CAPQuaM to develop this measure as part of the initial set of inpatient perinatal measures developed in the PQMP. Clinically, we have demonstrated that the temperature of low birthweight neonates is variable, and is highly consequential in terms of critical outcomes like survival and intraventricular hemorrhage. Institutional anecdotal evidence supports literature observations that thermal management can be managed and improved at the unit level with improved outcomes.

1c.4. Citations for data demonstrating high priority provided in 1a.3

1.Garrison FH. History of Pediatrics in Abt IA. Pediatrics. Volume I. WB Saunders. Philadelphia. 1923..
2.Holt LE. The Care of Premature and Delicate Infants in Holt LE. The Diseases of Infancy and Childhood, Second Edition Revised and Enlarged. D. Appleton. NY 1902
3.Baker, JP. The Incubator and the Medical Discovery of the Premature Infant. Journal of Perinatology 5:321-328. 2000.
4.Pierce RV. The People's Medical Adviser in Plain English. World's Dispensary Printing Office. Buffalo, NY 1875
5.Currier, A. Diseases of the Newborn. In Sajous CE. (editor). Annual of the Universal Medical Sciences. FA Davis, Philadelphia. 1891. Volume II, Section L.
6.Fischer L/ Diseases of Infancy and Childhood. FA Davis and Co. 1915.
7.Holt Jr. EL, McIntosh R. Holt's Diseases of Infants and Children, 11th Edition. Appleton. New York. 1940.
8.Silverman WA, Fertig JW, Berger AP. The influence of the thermal environment upon the survival of newly born premature babies. Pediatrics. 22:876-886. 1958.
9.Sinclair JC. Servo-control for maintaining abdominal skin temperature at 36C in low birth weight infants. Cochrane Database of Systematic Reviews 2007, Issue 4, Art. No. CD001074.
10.Watkinson M. Temperature Control of Premature Infants in the Delivery Room. Clinics in Perinatology 33:43-53.2006
11.Laptook AR, Salhab W, Bhaskar B, et al. Admission Temperature of Low Birth Weight Infants.; Predictors and Associated Morbidities. Pediatrics. 119:e643-649. 2007.
12.Howell EA, Stone J, Inamdar S, Matseoane S, Kleinman LC, Chassin MR. Approaching Guideline Recommended Care for Maternal-Infant Health: Clinical failures to use recommended antenatal corticosteroids, Maternal Child Health Journal, 2009 Jun 4. [Epub ahead of print]
13.Yazidi, GA., Boudes, E., Tan X., et al. Intraventricular hemorrhage in asphyziated newobords treated with hypothermia: a look into incidence, timing and risk factors. BMC Pediatr. 15. 2015. 106.

Additional citations:

Agency for Healthcare Research and Quality. National Quality Measures Clearinghouse http://qualitymeasures.ahrq.gov/tutorial/validity.aspx. Accessed February 6, 2012.
AHRQ, National Quality Measures Clearinghouse Glossary http://qualitymeasures.ahrq.gov/about/glossary.aspx. Accessed February 1, 2012.
CHIPRA, 2009. Children's Health Insurance Program Reauthorization Act of 2009. Pub. L. no. 111-3, 123 Stat. 8 (2009). http://frwebgate.access.gpo.gov/cgi-bin/getdoc.cgi?dbname=111_cong_public_laws&docid=f:publ003.111. Accessed February 6, 2012.
Cohn, V, 2001. News and numbers: A guide to reporting statistical concepts in health and other fields. Ames, I.A.: Wiley-Blackwell.
CMS. Centers for Medicare & Medicaid Services Quality Measurement and Health Assessment Group Glossary, as part of the Measures Management System Measure Development Overview

http://www.cms.gov/MMS/19_MeasuresManagementSystemBlueprint.asp#TopOfPage. Accessed February 6, 2012.
Donabedian, A. 1980. Methods for deriving criteria for assessing the quality of medical care. Medical Care Review 37(7):653–698.
Donabedian, A. 1988. The quality of care. How can it be assessed? Journal of the American Medical Association 260(12):1743–1748.
Eisenberg-Testimony on Health Care Quality. John Eisenberg, MD, Administrator, AHCPR, before the House Subcommittee on Health and the Environment, October 28, 1997. Agency for Health Care Policy and Research, Rockville, MD.
http://archive.ahrq.gov/news/test1028.htm. Accessed February 6, 2012.
IOM, 1990. Institute of Medicine, Medicare: A Strategy for Quality Assurance, Volume 1. National Academies Press, 1990.
IOM, 2001. Institute of Medicine, Crossing the Quality Chasm: A New Health System for the 21st Century. Washington, DC: National Academies Press; 2001.
IOM, 2006. Appendix E. Institute of Medicine, Committee on Redesigning Health Insurance Performance Measures, Payment, and Performance Improvement Programs. Performance Measurement: Accelerating Improvement (Pathways to Quality Health Care Series). Appendix E. The National Academies Press; 2006.
Kunkel, S., Rosenqvist, U., and Westerling, R. 2007. The structure of quality systems is important to the process and outcome, an empirical study of 386 hospital departments in Sweden. BMC Health Services Research Jul 9; 7:104.
http://www.ncbi.nlm.nih.gov/pubmed/17620113
National Committee on Quality Assurance's Practice Site Methodology. Definition of Practice Site adapted by the PQMP Result Aggregation Workgroup, 2012.
Patient Protection and Affordable Care Act of 2010. P.L. 111-148. Washington, DC. U.S. Government Printing Office, 2010.
http://housedocs.house.gov/energycommerce/ppacacon.pdf. Accessed February 6, 2012.
PQMP Result Aggregation Workgroup, 2012. AHRQ-CMS Pediatric Quality Measures Program, PQMP Performance Result Aggregation Framework; 2012.
The Free Dictionary by Farlex: http://www.thefreedictionary.com/test. Accessed February 6, 2012.
RTI. RTI International Staff. http://www.rti.org.
Wikipedia, based on The Standards for Educational and Psychological Testing, 1999. American Educational Research Association, Psychological Association, & National Council on Measurement in Education. Washington, DC: American Educational Research Association; 1999***. http://en.wikipedia.org/wiki/Standards_for_Educational_and_Psychological_Testing. Accessed February 6, 2012. ***A revised version is expected after 2012.

Other numbered citations
1. Hall SA, Kaufman JS, Ricketts TC. Defining urban and rural areas in U.S. epidemiologic studies. J Urban Health. 2006;83(2):162-75. PMCID: 2527174.
2. Hankins GD, Clark SL, Pacheco LD, O´Keeffe D, D´Alton M, Saade GR. Maternal mortality, near misses, and severe morbidity: lowering rates through designated levels of maternity care. Obstet Gynecol. 2012;120(4):929-34.
3. Berg CJ, Callaghan WM, Syverson C, Henderson Z. Pregnancy-related mortality in the United States, 1998 to 2005. Obstet Gynecol. 2010;116(6):1302-9.
4. Callaghan WM, Creanga AA, Kuklina EV. Severe maternal morbidity among delivery and postpartum hospitalizations in the United States. Obstet Gynecol. 2012;120(5):1029-36.
5. Howell EA, Stone J, Kleinman LC, Inamdar S, Matseoane S, Chassin MR. Approaching NIH guideline recommended care for maternal-infant health: clinical failures to use recommended antenatal corticosteroids. Matern Child Health J. 2010;14(3):430-6.
6. Howell EA, Holzman I, Kleinman LC, Wang J, Chassin MR. Surfactant use for premature infants with respiratory distress syndrome in three New York city hospitals: discordance of practice from a community clinician consensus standard. J Perinatol. 2010;30(9):590-5. PMCID: 2888640.
7. Howell EA, Hebert P, Chatterjee S, Kleinman LC, Chassin MR. Black/white differences in very low birth weight neonatal mortality rates among New York City hospitals. Pediatrics. 2008;121(3):e407-15.
8. Hoyert DL, Danel I, Tully P. Maternal mortality, United States and Canada, 1982-1997. Birth. 2000;27(1):4-11.
9. Berg CJ, Atrash HK, Koonin LM, Tucker M. Pregnancy-related mortality in the United States, 1987-1990. Obstet Gynecol. 1996;88(2):161-7.
10. Nannini A, Weiss J, Goldstein R, Fogerty S. Pregnancy-associated mortality at the end of the twentieth century: Massachusetts, 1990-1999. J Am Med Womens Assoc. 2002;57(3):140-3.

11. Commission TJ. Preventing Maternal Death: Sentinel Event Alert; 2010 January 26, 2010 Contract No.: Document Number|.
12. Levels of neonatal care. Pediatrics. 2012;130(3):587-97.
13. Collaborative CMQC. OB Hemorrhage Toolkit. 2010 [updated 2010; cited]; Available from: http://cmqcc.org/ob_hemorrhage.
14. DeFrances CJ, Cullen KA, Kozak LJ. National Hospital Discharge Survey: 2005 annual summary with detailed diagnosis and procedure data. Vital Health Stat 13. 2007(165):1-209.
15. Andrews R. The national hospital bill: the most expensive conditions by payer. HCUP Statistical Brief 59. Rockville, MD: Agency for Healthcare, Research and Quality; 2008.
16. Health NYSDo. Table 4: Live Birth Summary by Mother´s Race/Ethnicity, New York State 2011. 2013 [updated 2013; cited]; Available from: http://www.health.ny.gov/statistics/vital_statistics/2011/table04.htm.
17. Association AH. Chart Book: Prepared to Care the 24/7 Role of America´s Full-service Hospitals; 2005 Contract No.: Document Number|.
18. Health NYSDo. New York State Hospital Profile 2012 [updated 2012; cited]; Available from: http://hospitals.nyhealth.gov/index.php.
19. Committee Opinion no. 514: emergent therapy for acute-onset, severe hypertension with preeclampsia or eclampsia. Obstet Gynecol. 2011;118(6):1465-8.
20. Bucklin BA, Hawkins JL, Anderson JR, Ullrich FA. Obstetric anesthesia workforce survey: twenty-year update. Anesthesiology. 2005;103(3):645-53.
21. ACOG committee opinion No. 433: optimal goals for anesthesia care in obstetrics. Obstet Gynecol. 2009;113(5):1197-9.
22. Barbieri R, Camann WR. A pivotal partnership: obstetrics and anesthesiology [Editorial]. OB Management. 2002(March):8-9.
23. Phibbs CS, Bronstein JM, Buxton E, Phibbs RH. The effects of patient volume and level of care at the hospital of birth on neonatal mortality. JAMA. 1996;276(13):1054-9.
24. McCormick MC, Shapiro S, Starfield BH. The regionalization of perinatal services. Summary of the evaluation of a national demonstration program. JAMA. 1985;253(6):799-804.
25. Aday LA, Andersen R. A framework for the study of access to medical care. Health Serv Res. 1974;9(3):208-20. PMCID: 1071804.
26. Andersen RM, McCutcheon A, Aday LA, Chiu GY, Bell R. Exploring dimensions of access to medical care. Health Serv Res. 1983;18(1):49-74. PMCID: 1068709.
27. Kuhlthau KA. Measures of availability of health care services for children. Acad Pediatr. 2011;11(3 Suppl):S42-8.
28. ACOG Committee Opinion No. 433: Optimal Goals for Anesthesia Care in Obstetrics; 2009 Contract No.: Document Number|.
29. Geller SE, Rosenberg D, Cox SM, Brown ML, Simonson L, Driscoll CA, et al. The continuum of maternal morbidity and mortality: factors associated with severity. Am J Obstet Gynecol. 2004;191(3):939-44.
30. Pasupathy D, Wood AM, Pell JP, Fleming M, Smith GC. Time of birth and risk of neonatal death at term: retrospective cohort study. BMJ. 2010;341:c3498. PMCID: 2904877.
31. Martin JA, Hamilton, B.E., Ventura S.J., Osterman M, WIlson E.C., Mathews T.J. Births: Final Data for 2010. National Center for Health Statistics; August 28, 2012 Contract No.: Document Number|.
32. Clark SL, Belfort MA, Dildy GA, Herbst MA, Meyers JA, Hankins GD. Maternal death in the 21st century: causes, prevention, and relationship to cesarean delivery. Am J Obstet Gynecol. 2008;199(1):36 e1-5; discussion 91-2 e7-11.
33. Shearer EL. Cesarean section: medical benefits and costs. Soc Sci Med. 1993;37(10):1223-31.
34. Ehrenthal DB, Chichester ML, Cole OS, Jiang X. Maternal risk factors for peripartum transfusion. J Womens Health (Larchmt). 2012;21(7):792-7.

35. Rouse DJ, MacPherson C, Landon M, Varner MW, Leveno KJ, Moawad AH, et al. Blood transfusion and cesarean delivery. Obstet Gynecol. 2006;108(4):891-7.
36. Chung JH, Phibbs CS, Boscardin WJ, Kominski GF, Ortega AN, Needleman J. The effect of neonatal intensive care level and hospital volume on mortality of very low birth weight infants. Med Care. 2010;48(7):635-44.
37. Medicine Io. Crossing the Quality Chasm: The IOM Health Care Quality Initiative, Contract No.: Document Number|.
38. Kuklina EV, Whiteman MK, Hillis SD, Jamieson DJ, Meikle SF, Posner SF, et al. An enhanced method for identifying obstetric deliveries: implications for estimating maternal morbidity. Matern Child Health J. 2008;12(4):469-77.
39. Diamond CRK, Kohler SA. . Use of Paper Medical Records Versus Administrative Data for Measuring and Improving Health Care Quality: Are We Still Searching for a Gold Standard? Disease Management. 2001;4(3):121-30.
40. Hospitals AASo. Data Collection Methods. Health Forum, LLC; [cited]; Available from: http://www.ahadataviewer.com/about/data/.
41. Virnig BA, McBean M. Administrative data for public health surveillance and planning. Annu Rev Public Health. 2001;22:213-30.
42. Mangione-Smith R, DeCristofaro AH, Setodji CM, Keesey J, Klein DJ, Adams JL, et al. The quality of ambulatory care delivered to children in the United States. N Engl J Med. 2007;357(15):1515-23.
43. Rubio DM B-WM, Tebb SS, Lee ES, Rauch S. Objectifying content validity: Conducting a content validity study in social work research. Social Work Research. 2003;27(2):94-104.
44. Bennett KJ, Olatosi B. & Probst, J.C. Health Disparities: A rural-urban chartbook. Columbia, South Carolina: South Carolina Rural Health Research Center; 2008.
45. Hart G. Frontier/Remote Island, and Rural Literature Review2012: Available from: http://ruralhealth.und.edu/frontier/pdf/lit_review.pdf.

1c.5. **If a PRO-PM (e.g. HRQoL/functional status, symptom/burden, experience with care, health-related behaviors), provide evidence that the target population values the measured PRO and finds it meaningful.** *(Describe how and from whom their input was obtained.)*
n/a

2. Reliability and Validity—Scientific Acceptability of Measure Properties

Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. ***Measures must be judged to meet the subcriteria for both reliability and validity to pass this criterion and be evaluated against the remaining criteria.***

2a.1. Specifications The measure is well defined and precisely specified so it can be implemented consistently within and across organizations and allows for comparability. eMeasures should be specified in the Health Quality Measures Format (HQMF) and the Quality Data Model (QDM).

De.5. **Subject/Topic Area** *(check all the areas that apply):*
Perinatal and Reproductive Health, Perinatal and Reproductive Health : Newborn, Perinatal and Reproductive Health : Screening, Prevention

De.6. **Cross Cutting Areas** *(check all the areas that apply):*
Disparities, Functional Status, Prevention : Social Determinants, Safety, Safety : Complications

S.1. **Measure-specific Web Page** *(Provide a URL link to a web page specific for this measure that contains current detailed specifications including code lists, risk model details, and supplemental materials. Do not enter a URL linking to a home page or to general information.)*
n/a

S.2a. **If this is an eMeasure**, HQMF specifications must be attached. Attach the zipped output from the eMeasure authoring tool (MAT) - if the MAT was not used, contact staff. (Use the specification fields in this online form for the plain-language description of the specifications)
This is not an eMeasure **Attachment:**

S.2b. **Data Dictionary, Code Table, or Value Sets** *(and risk model codes and coefficients when applicable) must be attached. (Excel or csv file in the suggested format preferred - if not, contact staff)*
Attachment **Attachment:** Thermal_ICD_Conversion_v2.xlsx

S.3. **For endorsement maintenance**, please briefly describe any changes to the measure specifications since last endorsement date and explain the reasons.
n/a

S.4. **Numerator Statement** *(Brief, narrative description of the measure focus or what is being measured about the target population, i.e., cases from the target population with the target process, condition, event, or outcome)*
IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
The metric of interest is the temperature upon arrival to the Level 2 or higher nursery that is being assessed. This measure does not have the form of numerator and denominator. It is a distribution. We ask for reporting of the distribution in terms of five categories across the distribution, in terms of key moments in the distribution, and as a graphical presentation of the distribution. This is an information rich measure. Accountability entities may choose to use any of various components for their emphasis (alone or in combination), including percent "about right", mean or median temperatures, or value of the 10th or 25th percentiles, and the inter-percentile range.

There is an eligible population of newborns, which could be considered the denominator.
In lieu of a numerator, this measure reports the distribution of temperatures, using both numbers and a graph. In order to allow for reporting of key factors of interest to the accountability entity, this measure is specified to report that distribution in a variety of ways. This measure offers users (the accountability entity) the option to focus on one or more key substantive aspects of thermal outcomes in the defined population.

Data Elements:

-- Temperature to first decimal place
-- Units of temperature (Celsius, Fahrenheit). Those measured in Fahrenheit should be converted to Celsius. Celsius=(Fahrenheit less 32) times 5 divided by 9.
-- Time that temperature was measured
-- Time of arrival to the nursery (not time that admission was done)

State and County of residence OR zip code of mother
-- Optional: Method of temperature measurement (axillary, rectal, skin, tympanic)

S.5. **Time Period for Data** *(What is the time period in which data will be aggregated for the measure, e.g., 12 mo, 3 years, look back to August for flu vaccination? Note if there are different time periods for the numerator and denominator.)*
Typically reported for one year. No fundamental barrier to using shorter periods when volume is sufficient.

S.6. **Numerator Details** *(All information required to identify and calculate the cases from the target population with the target process, condition, event, or outcome such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
IF an OUTCOME MEASURE, describe how the observed outcome is identified/counted. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
No numerator exclusions. The distribution of temperature for all eligible newborns is to be reported.

The value to be reported is the first temperature upon arrival to the nursery. It should be reported in Celsius to one decimal place.

The distribution should be described as follows.

A. Categorical description

We specify five categories based upon the temperature value

- Category 1 “Cold”: All neonates with temperatures less than or equal to 34.5° Celsius
- Category 2 “Very Cool”: All neonates with temperatures greater than 34.5° Celsius and less than or equal to 35.5° Celsius
- Category 3 “Cool”: All neonates with temperatures greater than 35.5° Celsius and less than or equal to 36.5° Celsius
- Category 4 “About Right”: All neonates with temperatures greater than 36.5° Celsius and less than or equal to 37.5° Celsius
- Category 5: “Overly warm”: All neonates with temperatures greater than 37.5° Celsius

The percent in each category is reported to 1 decimal places and is calculated as the number of temperatures meeting the criteria for that category divided by the denominator, that value multiplied by 100.

While our expert panel recommended and we specify as optimal the broader description of categories across the spectrum as indicated above, we understand that at times an accountability entity may prefer to summarize these categories as a single number. In such cases we recommend adding the number in Category 1, Category 2, and Category 3 and dividing that sum by the denominator and multiplying that proportion by 100 to calculate the percent. We term this percent to represent the number that are “Too Cool”. This dichotomization lacks the granularity of the full categorization and as such may be less useful to drive improvement efforts than the full categorization. It also does not provide a check for gaming of the system by harming babies through overwarming.

B. Non-categorical description of the distribution

The non-categorical description of the distribution of the measure calls for the presentation of a cumulative distribution curve, for which the left axis is the percent of admissions with or lower than the temperature on the x-axis and the x-axis is temperature in Celsius running from 34 to 38 degrees with each tenth of a degree indicated by a tick mark. The Y-axis should have tick marks at every 5% with lines at every 20%.

This measure requests the reporting of the following:

- Descriptors of the Center of the Distribution (for sample size >5)

a. Mean
b. Median (50th percentile)
c. 25th percentile
d. 75th percentile

- Descriptors of Dispersion

a. Interquartile range (for sample size >5)
b. Standard deviation (for sample size >10)

- Descriptors of the Warm End of the Distribution

a.99th percentile (for sample size >33)
b.95th percentile (for sample size >20)
c.90th percentile (for sample size >10)

- Descriptors of the Cool End of the Distribution

a.1st percentile (for sample size >33)
b.5th percentile (for sample size >20)
c.10th percentile (for sample size >10)

The Appendix includes an exemplar presentation of the distribution aspect of the measure and an exemplar pie chart of the categorical presentation of the measure. It also illustrates how the graphics can be used to visually compare distributions.

Our expert panel recommended reporting of the full distribution and we so specify here. If an accountability entity wanted to summarize findings with fewer numbers, we would recommend reporting of either the 10th percentile and the 90-10 inter-percentile range (calculated as the value associated with the 90th percentile less the value associated with the 10th percentile), OR

the 25th percentile and the interquartile range (calculated as the value associated with the 75th percentile less the value associated with the 25th percentile). This is analogous to the reporting of a mean and standard deviation in other circumstances.

The reporting at the 10th percentile emphasizes awareness of the lower end of the spectrum towards those children at risk for being Cold, while using the 25th percentile emphasizes the middle half of the distribution. The 90th percentile (which can be calculated from the 10th percentile and the 90-10 inter-percentile range provides modest insight into the warming end of the distribution that may help to identify systematic overwarming. Consideration of the middle of the distribution with the 25th percentile and interquartile range is far less sensitive regarding systematic overwarming.

S.7. Denominator Statement *(Brief, narrative description of the target population being measured)*
All newborn infants born in a medical facility with birthweights less than 2,500 grams and admitted to a level 2 or higher nursery within 24 hours of life, other than those excluded.

S.8. Target Population Category *(Check all the populations for which the measure is specified and tested if any):*
Children's Health, Populations at Risk

S.9. Denominator Details *(All information required to identify and calculate the target population/denominator such as definitions, specific data collection items/responses , code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
Live-born neonates with birthweight of less than 2,500 grams (as identified from either the medical record or by ICD-10-CM Principal or Other Diagnosis Codes in Table 1 admitted to a Level 2 or higher nursery within 24 hours of birth. ICD-09-CM Codes are available in Table 1 for time frames that have not yet converted to ICD-10-CM.

Children identified as having received Level 2 care either via medical record review and/or via revenue code 172, 173, or 174 shall be eligible for the denominator.

Data Elements
- Infant birthweight or category
- Date of Birth
- Time of Birth
- Time of arrival to level 2 or higher nursery
- State and County of residence of mother or zip code

Table 1. Included Populations: ICD-10-CM Principal or Other Diagnosis Code\Low Birthweight Diagnostic Codes

P05.00	Newborn light for gestational age, unspecified weight
P05.01	Newborn light for gestational age, less than 500 grams
P05.02	Newborn light for gestational age, 500-749 grams
P05.03	Newborn light for gestational age, 750-999 grams
P05.04	Newborn light for gestational age, 1000-1249 grams
P05.05	Newborn light for gestational age, 1250-1499 grams
P05.06	Newborn light for gestational age, 1500-1749 grams
P05.07	Newborn light for gestational age, 1750-1999 grams
P05.08	Newborn light for gestational age, 2000-2499 grams
P05.10	Newborn small for gestational age, unspecified weight
P05.11	Newborn small for gestational age, less than 500 grams
P05.12	Newborn small for gestational age, 500-749 grams
P05.13	Newborn small for gestational age, 750-999 grams
P05.14	Newborn small for gestational age, 1000-1249 grams
P05.15	Newborn small for gestational age, 1250-1499 grams
P05.16	Newborn small for gestational age, 1500-1749 grams
P05.17	Newborn small for gestational age, 1750-1999 grams
P05.18	Newborn small for gestational age, 2000-2499 grams
P05.2	Newborn affected by fetal (intrauterine) malnutrition not light or small for gestational age
P05.9	Newborn affected by slow intrauterine growth, unspecified

P07.00 Extremely low birth weight newborn, unspecified weight
P07.01 Extremely low birth weight newborn, less than 500 grams
P07.02 Extremely low birth weight newborn, 500-749 grams
P07.03 Extremely low birth weight newborn, 750-999 grams
P07.10 Other low birth weight newborn, unspecified weight
P07.14 Other low birth weight newborn, 1000-1249 grams
P07.15 Other low birth weight newborn, 1250-1499 grams
P07.16 Other low birth weight newborn, 1500-1749 grams
P07.17 Other low birth weight newborn, 1750-1999 grams
P07.18 Other low birth weight newborn, 2000-2499 grams

For codes P05.00, P05.10, P05.2, P05.9, P07.00, P07.10 birthweights should be verified from the medical record prior to including in measure.

S.10. **Denominator Exclusions** *(Brief narrative description of exclusions from the target population)*
Neonates with anencephaly, who receive only comfort care in the Level 2 or higher nursery, or those who die or are placed intentionally on a pre-existing hypothermia protocol prior to the 15 minute after arrival specification time.

S.11. **Denominator Exclusion Details** *(All information required to identify and calculate exclusions from the denominator such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
• Neonates who do not survive until the time limit of the measure (15 minutes after arrival to the Level 2 or higher nursery)
• Neonates with anencephaly ICD-10-CM Q00.0
• Neonates not born in hospital/medical care setting
• Neonates with Comfort care (requires all of the features below):
o Died within 48 hours of birth AND received no respiratory support after arrival to the Level 2 or higher nursery other than blow by oxygen (i.e., did not receive CPAP, intubation, or CPR after arrival at Level 2 or higher nursery)
• Neonates for whom the hospital provides documentation that at the time of arrival to the NICU and before the temperature was taken the infant had been identified as meeting written institutional criteria for the initiation of therapeutic hypothermia and such therapy was begun or planned (optional exclusion).

S.12. **Stratification Details/Variables** *(All information required to stratify the measure results including the stratification variables, definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format with at S.2b)*
To avoid the potential for gaming by delay of temperature, the results are specified that they must be stratified in three ways:
- Main Stratum: Time between arrival at Level 2 or higher nursery is between 0 and 15 minutes.
- Delayed stratum: Time between arrival at Level 2 or higher nursery is more than 15 minutes.
- Other: Inadequate documentation to determine timing of temperature

To identify racial and ethnic disparities, the measure should be stratified by Race/Ethnicity using at least the following categories: Hispanic, Non Hispanic White, Non Hispanic Black, Asian/Pacific Islander, and Other non-Hispanic. When N <15 for Asian/Pacific Islander or Native American, these children may be aggregated with the Other category. Accountability entities may request further specificity of categories (e.g. Native American).

In addition to reporting overall temperatures, we suggest that when sample size is sufficient (see below) the accountability agency request stratification by birthweight as follows:
<= 999 grams.
1000 to 1499 grams
1500 to 2499 grams

Recommended stratifications for entities that care for newborns from multiple counties include:
Rurality/Urbanicity of home county of mother;
Level of Poverty of home county of mother.

Optional stratifications based upon optional data elements. These stratifications may be requested by the accountability agency.

The optional data elements for stratification are:

Five minute Apgar: Stratification categories are:
Apgar Category 1: Five minute Apgar from 0 through 5;
Apgar Category 2: Five minute Apgar greater than or equal to 6;

Insurance type and benefit category
Private/Commercial
Medicaid
Uninsured
Within The Private and Medicaid insurance categories, further stratification may be requested:
HMO, PPO, Medicaid Primary Care Management Plan, Fee for Service, Other

Location of birth
Operating Room (includes all sites typically used for c-section)
Labor and Delivery/Birthing Room
Other

Route of delivery
Vaginal
C-section

Admission Source
Inborn and admitted from site of delivery to special/intensive care nursery
Admitted/transferred from newborn nursery at same facility/campus prior to admission to special/intensive care nursery
Transferred/transported from another facility, where delivery occurred

S.13. Risk Adjustment Type (Select type. Provide specifications for risk stratification in S.12 and for statistical model in S.14-15)
Stratification by risk category/subgroup
If other:

S.14. Identify the statistical risk model method and variables *(Name the statistical method - e.g., logistic regression and list all the risk factor variables. Note - risk model development and testing should be addressed with measure testing under Scientific Acceptability)*
n/a

S.15. Detailed risk model specifications *(must be in attached data dictionary/code list Excel or csv file. Also indicate if available at measure-specific URL identified in S.1.)*
Note: Risk model details (including coefficients, equations, codes with descriptors, definitions), should be provided on a separate worksheet in the suggested format in the Excel or csv file with data dictionary/code lists at S.2b.

S.15a. Detailed risk model specifications (*if not provided in excel or csv file at S.2b*)
n/a

S.16. Type of score:
Continuous variable, e.g. average
If other:

S.17. Interpretation of Score *(Classifies interpretation of score according to whether better quality is associated with a higher score, a lower score, a score falling within a defined interval, or a passing score)*
Better quality = Score within a defined interval

S.18. Calculation Algorithm/Measure Logic *(Describe the calculation of the measure score as an ordered sequence of steps including identifying the target population; exclusions; cases meeting the target process, condition, event, or outcome; aggregating data; risk*

adjustment; etc.)
Step 1: Identify all live-born neonates with a birthweight less than 2,500 grams, using the aforementioned codes or recorded birthweights when practical.

Step 2: Identify all neonates from Step 1 who were admitted to Level 2 or higher nursery.

Step 3: Record required attributes:
i. ICD-9 comorbid diagnoses. Exclude those with anencephaly (ICD-9-CM 740xx).
ii. Infant birthweight or category
iii. Date of birth
iv. Time of birth
v. Date and Time of arrival to level 2 or higher nursery
vi. ZIP Code, State and County or equivalent area of Mother's residence and FIPS if available
vii. Race
viii. Ethnicity
ix. Evidence child received comfort care only (when appropriate) and exclude if so
x. Documentation child was eligible for and received therapeutic hypothermia (when appropriate) and exclude if so
xi. If transferred, if there is documentation that neonate was not born in any medical facility and exclude if so.

Step 4: Record the following optional data elements as specified by the accountability agency:
i. Five minute Apgar score
ii. Insurance type (Medicaid, Commercial, Uninsured)
iii. Benefit category (HMO, PPO, Medicaid Primary Care Management Plan, Fee for Service, Other)
iv. Location of birth
v. Route of delivery
vi. Date, Time, Location and Value in Celsius of first Temperature taken (any location)
vii. Route of temperature taken upon admission to Level 2 or higher Nursery
viii. Route and location of first temperature taken if not in Level 2 or higher nursery
ix. If child was admitted to the Level 2 or higher nursery from regular newborn care
x. If child was inborn or transferred to Level 2 or higher nursery from another facility.

Step 5: Identify and record the following:
i. Time of first documented temperature taken in the nursery (arrival temperature)
ii. Value of first temperature taken in the nursery
iii. Units that temperature was recorded in. If in Fahrenheit, calculate Celsius as C = (F-32)*5/9.
iv. Infant age at time of arrival temperature

Step 6: If infant's age is > 75 minutes at the time of the initial temperature, record the following as the alternate temperature:
i. Time of last temperature taken in the unit where the infant was delivered
ii. Value of that temperature
iii. Units that temperature was recorded in. If in Fahrenheit, calculate Celsius as C = (F-32)*5/9.
iv. Infant age at time of temperature (may be calculated from time of birth and time of temperature)

If infant's age at time of arrival temperature is > 75 minutes AND infant was admitted directly to the Level 2 or higher nursery without transport from another institution or transfer from the normal newborn nursery, report the lower of the arrival and the alternate temperature.

Step 7: Identify which category the reported temperature should be assigned. Each is determined by the number of children whose reported temperature falls within the criteria for that category.
- Category 1 "Cold": All neonates with temperatures less than or equal to 34.5° Celsius
- Category 2 "Very Cool": All neonates with temperatures greater than 34.5° Celsius and less than or equal to 35.5° Celsius
- Category 3 "Cool": All neonates with temperatures greater than 35.5° Celsius and less than or equal to 36.5° Celsius
- Category 4 "About Right": All neonates with temperatures greater than 36.5° Celsius and less than or equal to 37.5° Celsius
- Category 5: "Overly warm": All neonates with temperatures greater than 37.5° Celsius

Step 8: Associate each admission with one of three timing strata.
- Main Stratum: Time between arrival at Level 2 or higher nursery and first temperature being taken in the Level 2 or higher nursery

is between 0 and 15 minutes.
- Delayed stratum: Time between arrival at Level 2 or higher nursery and first temperature being taken in the Level 2 or higher nursery is more than 15 minutes.
- Other: Inadequate documentation to determine timing of temperature in relationship to arrival in the Level 2 or higher nursery

Step 9: Calculate the percent of neonates who are in each stratum as
[100*number of children in each stratum] / [total number of infants eligible for the measure]. Percent should be reported to two decimal places. Minimum sample size for reporting the categorical findings for the nursery is N=20; reporting should be further stratified by the application of stratification variables as described below to both the numerator and the denominator. Reporting strata with denominator samples less than N=15 should not be reported.

Step 10: (Optional to be decided by accountability entity) Calculate 90% confidence intervals around each reported percentage as 1.68*[square root of the [proportion in the confidence interval*(one minus that percentage) / the eligible N].

Step 11: Describe the distribution of the temperatures. Minimum sample sizes for several moments are shown. Report all temperatures with their associated variables, N, and the following descriptors.
This measure requests the reporting of the following:

i.Descriptors of the Center of the Distribution (for sample size >5)
  a.Mean
  b.Median (50th percentile)
  c.25th percentile
  d.75th percentile

ii.Descriptors of Dispersion
  a.Interquartile range (for sample size >5)
  b.Standard deviation (for sample size >10)

iii.Descriptors of the Warm End of the Distribution
  a.99th percentile (for sample size >33)
  b.95th percentile (for sample size >20)
  c.90th percentile (for sample size >10)

iv.Descriptors of the Cool End of the Distribution
  a.1st percentile (for sample size >33)
  b.5th percentile (for sample size >20)
  c.10th percentile (for sample size >10)

Step 12: Using eligible births and qualified temperatures, repeat steps 9 and 10 to report for each stratification category listed below, using the following data elements:
i.Birthweight (3 birthweight categories: <999 grams; 1,000-1,499 grams; 1,500-2,499 grams)
a.Within each birthweight category, stratify by race/ethnicity. Include at least the following categories: Hispanic, Non-Hispanic Black, Non-Hispanic White; Non-Hispanic Asian/Pacific Islander, Other Non-Hispanic
Report strata for distributions when the number of infants in that category is at least 5. Report strata for categories when the sample size in that category is at least 15.

For optional stratifications as requested by the Accountability Entity, the following rules are hierarchical. In other words, categorization stops when the rules are applied in the given order and an infant meets the specified category. Hence it is critical that these rules are applied in order. Stop when categorized.

i.Categorize location of delivery as birthing room if:
1.Location was identified as delivery room on the labor and delivery suite but was not an operating room OR
2.Location was identified as a birthing room or equivalent OR
3.Infant was a vaginal delivery other than a multiple gestation AND Operating Room or equivalent (C-section room would be an example of an equivalent to an operating room) is not specified as location.

ii.Otherwise, categorize location as operating room if:

1.Location was identified as an operating room or equivalent, OR
2.If neonate was delivered by C-section OR
3.If infant was a multiple gestation (and location is unspecified) OR
4.If location is identified as Emergency Department OR other

b.5 minute Apgar score (Apgar of 5 or less versus 6 or more)

c.Benefit Category: HMO vs PPO vs FFS vs PCCM vs Other

d.Urban Influence Code. Identify the Urban Influence Code (1) or UIC. (2013 urban influence codes available at: http://www.ers.usda.gov/data-products/urban-influence- codes.aspx#.UZUvG2cVoj8). Use mother's place of residence to determine UIC. State and County names can be linked or looked up directly or ZIP codes can be linked to County indirectly, using the Missouri Census Data Center (http://mcdc.missouri.edu/). These data will link to County or County equivalents as used in various states.

e.Identify the Level of Poverty in the mother's county of residence. The percent of all residents in poverty by county or county equivalent are available from the US Department of Agriculture at http://www.ers.usda.gov/data-products/county-level-data-sets/download- data.aspx. Our stratification standards are based on 2011 US population data that we have analyzed with SAS 9.3. Using Mother's state and county of residence (or equivalent) or FIPS code, use the variable PCTPOVALL_2011 to categorize into one of 5 Strata:
i.Lowest Quartile of Poverty if percent in poverty is <12%
ii.Second Quartile of Poverty if percent in poverty is >12.5% and <16.5%
iii.Third Quartile of poverty if percent in poverty is >16.5% and <20.7%
iv.First Upper Quartile (75th-90th) if percent in poverty is >20.7% and <25.7%
v.Second Upper Quartile (>90th percentile) if percent in poverty exceeds 25.7%

ii.Repeat stratifications within birthweight categories(report for all strata for which denominator >=15)

S.19. **Calculation Algorithm/Measure Logic Diagram URL or Attachment** *(You also may provide a diagram of the Calculation Algorithm/Measure Logic described above at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
Available in attached appendix at A.1

S.20. **Sampling** *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*
IF a PRO-PM, identify whether (and how) proxy responses are allowed.
n/a
S.21. **Survey/Patient-reported data** *(If measure is based on a survey, provide instructions for conducting the survey and guidance on minimum response rate.)*
IF a PRO-PM, specify calculation of response rates to be reported with performance measure results.
n/a
S.22. **Missing data** (specify how missing data are handled, e.g., imputation, delete case.)
Required for Composites and PRO-PMs.
Delete cases from stratifications when data elements required to assign the strata are missing. Note and report the number of otherwise eligible infants omitted from that stratification.

Given the importance of temperature measurement in this population, there should be no missing temperatures or body weights. When calculating the distribution of temperatures, all missing temperatures should be considered as if the recorded temperature was 34 degrees. For the stratifications, the infants should be omitted and the number omitted reported.

S.23. **Data Source** *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*
If other, please describe in S.24.
Administrative claims, Electronic Clinical Data : Electronic Health Record, Other, Paper Medical Records

S.24. **Data Source or Collection Instrument** *(Identify the specific data source/data collection instrument e.g. name of database, clinical registry, collection instrument, etc.)*
IF a PRO-PM, identify the specific PROM(s); and standard methods, modes, and languages of administration.
1.Medical record (paper or electronic), may be utilized to identify the following data elements:

a.Date and time of birth
b.Date and time of arrival to a Level 2 or higher nursery
c.Date and time of first temperature upon arrival to that nursery
d.Temperature and units of measurement
e.Race/ethnicity (preferred data source)
f.Mother's State and County of Residence and or ZIP code (preferred data source)
g.Born in medical facility or transferred in (preferred data source)
h.5 minute Apgar score
i.Birthweight (preferred data source)
j.Documentation if child met local criteria for hypothermia and time so identified
k.Documentation if hypothermia was planned or initiated before temperature taken
l.Insurance type (optional data source)

2.Administrative data with revenue, billing and diagnosis codes, utilized to identify:
a.ICD-10 codes (to identify low birthweight infants and presence of anencephaly)
b.Revenue codes indicating care in Level 2, 3, or 4 nursery (172, 173, 174)
c.OPTIONAL source for:
i.Date of birth
ii.Race/ethnicity
iii.Home ZIP code
iv.Whether child was inborn or transferred in
v.Birthweight range
vi.Insurance type and benefit plan (preferred data source)

Data elements necessary for this measure include the following:
•Date/time of delivery,
•Date/time/value of first temperature after delivery and through admission to a Level 2 or higher nursery,
•Infant characteristics (birthweight, Apgar),
•Delivery characteristics (e.g., location of delivery, nursery level, delivery type), and
•Demographics (e.g., race, ethnicity, insurance, ZIP code).

Data elements necessary for this measure include: date/time of delivery, date/time/value of temperature after delivery and through the admission temperature at the Level 2 or higher nursery, infant characteristics (e.g., birthweight, Apgar), delivery characteristics (e.g., location of delivery, nursery level, delivery type), and demographics (race, ethnicity, insurance, ZIP code).

We also developed and assessed a web portal that allowed for clinical units to report medical record number, birthweight, arrival time, time of first temperature, and value of first temperature in real or near real time. More information about the portal is available upon request.

S.25. **Data Source or Collection Instrument** *(available at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
No data collection instrument provided

S.26. **Level of Analysis** *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*
Facility, Health Plan, Integrated Delivery System, Population : Community, Population : County or City, Population : Regional, Population : State

S.27. **Care Setting** *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*
Hospital/Acute Care Facility, Other
If other: Birthing Center

S.28. **COMPOSITE Performance Measure** - Additional Specifications *(Use this section as needed for aggregation and weighting rules, or calculation of individual performance measures if not individually endorsed.)*
n/a

2a. Reliability – See attached Measure Testing Submission Form
2b. Validity – See attached Measure Testing Submission Form
Thermal_Condition_nqf_testing_attachment_Submission_03_22_16.docx

NATIONAL QUALITY FORUM—Measure Testing (subcriteria 2a2, 2b2-2b7)

Measure Number (*if previously endorsed*)**:** Click here to enter NQF number

Measure Title: Thermal Condition of Low Birthweight Neonates Admitted to Level 2 or Higher Nurseries in the First 24 Hours of Life: A PQMP Measure

Date of Submission: 3/22/2016

Type of Measure:

☐ Composite – ***STOP – use composite testing form***	☒ Outcome (*including PRO-PM*)
☐ Cost/resource	☐ Process
☐ Efficiency	☐ Structure

Instructions

- Measures must be tested for all the data sources and levels of analyses that are specified. ***If there is more than one set of data specifications or more than one level of analysis, contact NQF staff*** about how to present all the testing information in one form.
- **For all measures, sections 1, 2a2, 2b2, 2b3, and 2b5 must be completed.**
- **For outcome and resource use measures**, section **2b4** also must be completed.
- If specified for **multiple data sources/sets of specificaitons** (e.g., claims and EHRs), section **2b6** also must be completed.
- Respond to all questions as instructed with answers immediately following the question. All information on testing to demonstrate meeting the subcriteria for reliability (2a2) and validity (2b2-2b6) must be in this form. An appendix for *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 20 pages (*incuding questions/instructions;* minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.
- For information on the most updated guidance on how to address sociodemographic variables and testing in this form refer to the release notes for version 6.6 of the Measure Testing Attachment.

Note: The information provided in this form is intended to aid the Steering Committee and other stakeholders in understanding to what degree the testing results for this measure meet NQF's evaluation criteria for testing.

2a2. Reliability testing [10] demonstrates the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise. For **PRO-PMs and composite performance measures**, reliability should be demonstrated for the computed performance score.

2b2. Validity testing [11] demonstrates that the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality. For **PRO-PMs and composite performance measures**, validity should be demonstrated for the computed performance score.

2b3. Exclusions are supported by the clinical evidence; otherwise, they are supported by evidence of sufficient frequency of occurrence so that results are distorted without the exclusion; [12]
AND

If patient preference (e.g., informed decisionmaking) is a basis for exclusion, there must be evidence that the exclusion impacts performance on the measure; in such cases, the measure must be specified so that the information about patient preference and the effect on the measure is transparent (e.g., numerator category computed separately, denominator exclusion category computed separately). [13]

2b4. For outcome measures and other measures when indicated (e.g., resource use):

- **an evidence-based risk-adjustment strategy** (e.g., risk models, risk stratification) is specified; is based on patient factors (including clinical and sociodemographic factors) that influence the measured outcome and are present at start of care; [14,15] and has demonstrated adequate discrimination and calibration

OR

- rationale/data support no risk adjustment/ stratification.

2b5. Data analysis of computed measure scores demonstrates that methods for scoring and analysis of the specified measure allow for **identification of statistically significant and practically/clinically meaningful [16] differences in performance**;

OR

there is evidence of overall less-than-optimal performance.

2b6. If multiple data sources/methods are specified, there is demonstration they produce comparable results.

2b7. For **eMeasures, composites, and PRO-PMs** (or other measures susceptible to missing data), analyses identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias.

Notes

10. Reliability testing applies to both the data elements and computed measure score. Examples of reliability testing for data elements include, but are not limited to: inter-rater/abstractor or intra-rater/abstractor studies; internal consistency for multi-item scales; test-retest for survey items. Reliability testing of the measure score addresses precision of measurement (e.g., signal-to-noise).

11. Validity testing applies to both the data elements and computed measure score. Validity testing of data elements typically analyzes agreement with another authoritative source of the same information. Examples of validity testing of the measure score include, but are not limited to: testing hypotheses that the measures scores indicate quality of care, e.g., measure scores are different for groups known to have differences in quality assessed by another valid quality measure or method; correlation of measure scores with another valid indicator of quality for the specific topic; or relationship to conceptually related measures (e.g., scores on process measures to scores on outcome measures). Face validity of the measure score as a quality indicator may be adequate if accomplished through a systematic and transparent process, by identified experts, and explicitly addresses whether performance scores resulting from the measure as specified can be used to distinguish good from poor quality.

12. Examples of evidence that an exclusion distorts measure results include, but are not limited to: frequency of occurrence, variability of exclusions across providers, and sensitivity analyses with and without the exclusion.

13. Patient preference is not a clinical exception to eligibility and can be influenced by provider interventions.

14. Risk factors that influence outcomes should not be specified as exclusions

15. With large enough sample sizes, small differences that are statistically significant may or may not be practically or clinically meaningful. The substantive question may be, for example, whether a statistically significant difference of one percentage point in the percentage of patients who received smoking cessation counseling (e.g., 74 percent v. 75 percent) is clinically meaningful; or whether a statistically significant difference of $25 in cost for an episode of care (e.g., $5,000 v. $5,025) is practically meaningful. Measures with overall less-than-optimal performance may not demonstrate much variability across providers.

1. DATA/SAMPLE USED FOR ALL TESTING OF THIS MEASURE

Often the same data are used for all aspects of measure testing. In an effort to eliminate duplication, the first five questions apply to all measure testing. If there are differences by aspect of testing,(e.g., reliability vs. validity) be sure to indicate the specific differences in question 1.7.

1.1. What type of data was used for testing? (*Check all the sources of data identified in the measure specifications and data used for testing the measure. Testing must be provided for all the sources of data specified and intended for measure implementation.* ***If different data sources are used for the numerator and denominator, indicate N [numerator] or D [denominator] after the checkbox.***)

Measure Specified to Use Data From: (*must be consistent with data sources entered in S.23*)	**Measure Tested with Data From:**
☒ abstracted from paper record	☒ abstracted from paper record
☒ administrative claims	☒ administrative claims
☐ clinical database/registry	☐ clinical database/registry
☒ abstracted from electronic health record	☒ abstracted from electronic health record
☐ eMeasure (HQMF) implemented in EHRs	☐ eMeasure (HQMF) implemented in EHRs
☒ other: Survey of hospital's quality team, state repository of data	☒ other: Survey of hospital's quality team, state repository of data

1.2. If an existing dataset was used, identify the specific dataset (*the dataset used for testing must be consistent with the measure specifications for target population and healthcare entities being measured; e.g., Medicare Part A claims, Medicaid claims, other commercial insurance, nursing home MDS, home health OASIS, clinical registry*).

NY State Department of Health Neonatal Database

1.3. What are the dates of the data used in testing? January 1, 1999 – December 31, 2002

1.4. What levels of analysis were tested? (*testing must be provided for all the levels specified and intended for measure implementation, e.g., individual clinician, hospital, health plan*)

Measure Specified to Measure Performance of: (*must be consistent with levels entered in item S.26*)	**Measure Tested at Level of:**
☐ individual clinician	☐ individual clinician
☐ group/practice	☐ group/practice
☒ hospital/facility/agency	☒ hospital/facility/agency
☒ health plan	☒ health plan
☒ other: integrated delivery system, population, state, region, county	☒ other: integrated delivery system, population, state, region, county

1.5. How many and which measured entities were included in the testing and analysis (by level of analysis and data source)? (*identify the number and descriptive characteristics of measured entities included in the analysis (e.g., size, location, type); if a sample was used, describe how entities were selected for inclusion in the sample*)

Testing included work done at 3 hospitals in New York City via chart abstraction (~700 infants included in analysis) and using a New York State DoH database that includes nearly 90% of all hospitals in the state that have Level 2 or higher nurseries and included a population of 7553 neonates from 61 nurseries across the state.

1.6. How many and which patients were included in the testing and analysis (by level of analysis and data source)? (*identify the number and descriptive characteristics of patients included in the analysis (e.g., age, sex, race, diagnosis); if a sample was used, describe how patients were selected for inclusion in the sample*)

All 7,553 infants under 2500 grams of birth weight admitted to one of 61 Level 2 or higher nurseries in NY State that submitted data to the data base.

Another 700+ infants under 2500 g born in one of 3 New York City hospitals in a distinct time frame and who had their charts reviewed.

1.7. If there are differences in the data or sample used for different aspects of testing (e.g., reliability, validity, exclusions, risk adjustment), identify how the data or sample are different for each aspect of testing reported below.

The analysis of the ~700 infants included regression analyses to look at the relationship of temperature with outcomes, including after adjusting for racial characteristics, Apgar scores, and other key variables. We also found racial differences in thermal and in clinical outcomes, such as death and brain hemorrhage.

The 7,553 infants were analyzed to show differences in performance from nursery to nursery, to describe the distribution of outcomes within and across populations, as well as to confirm our other findings that temperature did not function simply as a threshold phenomenon.

1.8 What were the patient-level sociodemographic (SDS) variables that were available and analyzed in the data or sample used? For example, patient-reported data (e.g., income, education, language), proxy variables when SDS data are not collected from each patient (e.g. census tract), or patient community characteristics (e.g. percent vacant housing, crime rate).

Race/ethnicity, and insurance status were available and analyzed.

2a2. RELIABILITY TESTING

Note*: If accuracy/correctness (validity) of data elements was empirically tested, separate reliability testing of data elements is not required – in 2a2.1 check critical data elements; in 2a2.2 enter "see section 2b2 for validity testing of data elements"; and skip 2a2.3 and 2a2.4.*

2a2.1. What level of reliability testing was conducted? (*may be one or both levels*)

☒ **Critical data elements used in the measure** (*e.g., inter-abstractor reliability; data element reliability must address ALL critical data elements*)

☐ **Performance measure score** (e.g., *signal-to-noise analysis*)

2a2.2. For each level checked above, describe the method of reliability testing and what it tests (*describe the steps—do not just name a method; what type of error does it test; what statistical analysis was used*)

The basis for the scientific soundness of this measure lies in the use of a hybrid of administrative/encounter and medical records data. Though they have their limitations, these data types have been shown in multiple studies to be a reliable source of information for reporting organizational performance. One such study found that quality measures that could be calculated using administrative data showed higher rates of performance than indicated by a review of the medical record alone, and that claims data is more accurate for identifying services with a high likelihood of documentation due to reimbursement (1).

This data relies on information in the medical record for the key variables, temperature, time of birth, and time of arrival in the nursery. Survey of quality offices across a variety of hospitals was performed by the Joint Commission under contract to CAPQuaM and this survey confirmed the availability, the feasibility, and confidence in the required data elements. As a vital sign of longstanding, temperature is familiar and well recorded. Cited guidelines (e.g. WHO, WAHT) confirm that route of measurement is not critical in this population premature infants and that commonly used methods to quantify temperature are sufficiently reliable. The most recommended routes for assessing temperature are axillary

and rectal, which are each reliable and valid when assessed in the eligible population when compared to one another. The landmark article which compared rectal and axillary temperatures in term and pre-term infants and found no meaningful differences in 99 term and 44 preterm infants, across a range of gestational ages.(2) Mariamma et al, specified that simultaneous measurement of rectum, femoral, and axillary temperatures in 99 infants were not different. (3) Moen et al studies 25 preterm infants with 12 simultaneous measurements over a two day period and found that axillary and rectal temperatures were sufficiently similar that one could be used for the other (4) Jirapaet and Jirapaet confirmed that rectal and axillary findings were similar in 57 term and 52 preterm infants and that abdominal skin temperature was similar within 0.2-0.3 degrees of axillary and rectal temperatures. (5)

A feasibility study designed to determine the ability and ease of collecting related data showed that date and time are self-evident and that there is mild but manageable variation in how time is reported. This should not impair the calculation of a neonate's age or the relationship of the time of measurement to the time of birth or to the time of arrival to the NICU as is required in our measure set. The underlying construct for temperature is the core body temperature of the neonate. For neonates of various sizes and gestational age, the optimal approach to measuring the temperature may vary. Measurement approaches that are understood to be valid may include rectal temperatures, tympanic temperatures, axillary temperatures, and when appropriately shielded from a radiant heat source, skin temperatures.

We understand that it would be a barrier to the wide adoption of this measure were we to specify changes to institutional standards of care regarding how to measure and record the temperature of low birthweight infants or to establish requirements for measurement given the current evidence in the literature. Therefore we do not offer such specification. Instead we ask that reporting agencies record and share the data regarding how each temperature was assessed so that the agencies receiving the data may use that information should they wish to do so.

The reliability of modern methods for assessing temperature is very high.

Robyn Knobel, RN, PhD was one of the experts on our panel; her doctoral dissertation documented high inter-rater reliability of nurses assessing infant temperature (kappa 0.96). The reliability of temperature extraction from medical records of sick neonates was assessed in the context of assessing the reliability of the Score of Acute Neonatal Physiology: Kappa for temperature was found to be 0.81 (95% CI: 0.71, 0.90). (6-7)

References

1. Fowles, J. B., Fowler, E. J., & Craft, C. (1998). Validation of claims diagnoses and self-reported conditions compared with medical records for selected chronic diseases. Journal of Ambulatory Care Management, 21(1), 24-34
2. *Mayfield SR, Bhatia J, Nakamura KT, Rios GR, Bell EF. Temperature measurement in term and preterm neonates. J Pediatr. 1984;104(2):271–275*,
3. Kunnel, M., O'brien, C., Munro, BH., Medoff-Cooper, B. (1988). *Comparisons of Rectal, Femoral, Axillary, and Skin-to-Mattress Temperatures in Stable Neonates*. Nursing Research (37(3).
4. Moen JE, Chapman S, Sheehan A, Carter P. Axillary versus rectal temperatures in preterm infants under radiant warmers. J Obstet Gynecol Neonatal Nurs. 1987;16(5):348–352.
5. Jirapaet and Jirapaet. (2000). Comparisons of tympanic membrane, abdominal skin, axillary, and rectal temperature measurements in term and preterm neonates. Nursing & Health Sciences 2(1):1-8.

6. Knobel, RN. Physiological effects of thermoregulation in transitional ELBW infants. 2006 https://cdr.lib.unc.edu/indexablecontent/uuid:1a02d564-0ae8-4c9f-b387-ce138a6f33cf (last accessed 3.19.2016)
7. Sutton L, Bajuk B, Berry G, Sayer GP, Eagles BL, Henderson–Smart DJ. Reliability of the SNAP (score of neonatal acute physiology) data collection in mechanically ventilated term babies in New South Wales, Australia. Acta Pediatr 91: 424-429. 2002

2a2.3. For each level of testing checked above, what were the statistical results from reliability testing? (e.*g., percent agreement and kappa for the critical data elements; distribution of reliability statistics from a signal-to-noise analysis*)

Validity testing was performed at the data element level for both the numerator and the denominator.

Our testing (using Mount Sinai data) of ICD-9 codes as a way to identify low birthweight infants found that 99 infants out of 677 who were identified with our ICD-9 specifications listed in Table 1, Section I, had birthweights of over 2,500. The ICD9 codes for this cohort that were 2,500 grams or above is listed in Table 2.

Table 2

ICD	N	Percent	Cumulative %
765.10	42	42.4%	42.4%
764.00	37	37.4%	79.8%
764.10	6	6.1%	85.9%
765.18	4	4.0%	89.9%
764.90	3	3.0%	92.9%
764.90,764.00	3	3.0%	96.0%
765.18,765.10	2	2.0%	98.0%
764.00,764.90	1	1.0%	99.0%
764.08	1	1.0%	100.0%
TOTAL	99		

Of the 99 infants, 5 had recorded birthweights of 2,500 grams, consistent with the ICD-9 codes used. We have indicated in our specifications that the various ICD-9 codes, such as 764.00, 764.10, and 765.10 that represent poor fetal growth without a specified weight need to have their eligibility for the measure confirmed with an actual birthweight.

2a2.4 What is your interpretation of the results in terms of demonstrating reliability? (i.*e., what do the results mean and what are the norms for the test conducted?*)

We are confident that the elements and the resulting measure are reliable. Not only does the evidence support this, but a conceptual model confirms that quantitative and concrete data elements, such as temperature, ought to be among the most reliable data when abstracted from a chart.

- Aaronson Ls, Burman, ME. (1994). Use of health records in research: reliability and validity issues. Res Nurs Health. Feb; 17 (1): 67-73.

2b2. VALIDITY TESTING

2b2.1. What level of validity testing was conducted? (*may be one or both levels*)

☒ **Critical data elements** (*data element validity must address ALL critical data elements*)

☒ **Performance measure score**

☐ **Empirical validity testing**

☐ **Systematic assessment of face validity of performance measure score as an indicator** of quality or resource use (*i.e., is an accurate reflection of performance on quality or resource use and can distinguish good from poor performance*)

2b2.2. For each level of testing checked above, describe the method of validity testing and what it tests (*describe the steps—do not just name a method; what was tested, e.g., accuracy of data elements compared to authoritative source, relationship to another measure as expected; what statistical analysis was used)*

The reliability section above contains some information related to validity as well. Modern instruments to assess temperature in the newborn are both reliable and valid and the literature base for both are the same. As noted above, guidelines allow for a variety of methods of assessment and depend upon the temperature that is assessed and recorded.

The use of electronically available administrative data in healthcare research and assessment is becoming increasingly common. Most databases contain consistent elements, are available in a timely manner, provide information about large numbers of individuals, and are relatively inexpensive to obtain and use. Validity has been established, and its strengths and weaknesses relative to data abstracted from medical records and obtained via survey have been documented (1) Administrative data are supported, if not encouraged by federal agencies, including NIH, AHRQ, HCFA, and the VA. This measure calls for the use administrative data to identify the universe of low birthweight infants. See reliability section above for our experience regarding validity.

The use of Expert Panels has been demonstrated to be useful in measure development and health care evaluation, including for children (2). Practitioners have been identified as a resource for researchers in developing and revising measures, since they are on the frontlines working with the populations who often become research participants. Involving practitioners can assist researchers in the creation of measures that are appropriate and easily administered.

The validity of our work has benefited from our use of a formal method, a pragmatic adaptation of the CAPQuaM 360° method. Our work represents a systematic assessment of the face and construct validity of our measure.

The 360° method (table 1) is highly engaged with collaborators, partners, and the literature. The explicit criteria that we use were developed using a slightly modified version of the RAND/UCLA Appropriateness Method that maintained the key aspects of that approach, including a detailed literature review, a multidisciplinary and geographically diverse expert panel comprised of both clinicians and researchers, and the two Round modified Delphi Process. It seeks to target relevant information and perspective and to have measures emerge from the process. The potential measures are then tested to the extent that time and resources permit. In developing the perinatal measures we incorporate:

- A high level of engagement with partnered institutions and senior advisors that bring into the process a wide diversity of stakeholders;
- A detailed literature review that is updated and supplemented as needed;
- Interviews with clinicians;
- The CAPQuaM scientific team (professionals qualified in neonatology, pediatrics, obstetrics and gynecology, epidemiology, quality measurement and improvement, patient safety, and public health);

- A geographically diverse, multidisciplinary Expert Panel who participated in a two-round RAND/UCLA modified Delphi process, with enhanced follow-up;
- Development of a Boundary Guideline that takes a multi-vectorial approach to incorporate simultaneously a variety of gradients, including gradients of importance, relevance, and certainty, as appropriate to the construct being represented;
- Specification and review of measures and approaches to measurement by stakeholders and experts; and
- Testing and assessment of measure performance to the extent feasible given resources and available time.

Table 1. 360 Degree Pediatric Quality Measure Development: Overview

Stage	Phase	Innovation	Product(s)
1. Clinical Criteria Development	a. Input Development	1. Focus groups of caregivers of children with asthma who have used the ED 2. Interviews with front line clinicians: primary care, asthma docs, and ED docs	1. Literature review 2. Clinician interviews (Family physicians, Obstetricians, neonatologists, pediatric hospitalists, general pediatricians)
	b. RAND/UCLA 2 Round Modified Delphi Process	1. Inclusion of consumer perspectives as a key input; 2. Use of this method to identify appropriateness criteria in national performance measure development;	1. Explicit criteria that rank a comprehensive and mutually exclusive set of clinically detailed scenarios;
2. Boundary Guideline Development	Criteria Enhancement	1. Iterative process to enhance reliability and internal consistency of the explicit criteria set with a goal of outlining three boundary spaces	1. Internally consistent set of explicit criteria that are stable in their representation of the expert panel perspective. "Enhanced criteria"
	Guideline Articulation	1. Stakeholder (including experts, users, clinicians, consumers and others) informed review of the enhanced criteria. 2. Definition of zones of potential overuse, potential underuse, and professional interaction and decision-making based upon the explicit criteria 3. Stakeholder valuations of potential deviations from guideline 4. Boundary Guideline	1. Boundary Guideline 2. Prioritization list
3. Creation of Measure	Specification	1. Translation of guideline into specification of necessary data 2. Iterative process to define optimally efficient sources of data to allow for measurement and stratification	1. Initial specification of measure
	Fielding and testing of measure	1. Measure testing	1. Functional experience and practical understanding of measure, its scoring, variability, and interpretation

Our feasibility work which included surveys of hospital quality measurement personnel confirms that the time that the temperature is assessed, rather than simply the time that it is documented, is recorded in the medical record, generally an EMR. This is a critical aspect of the validity of time data.

Our underlying construct is core body temperature. Modern temperatures are valid and precise. The core body temperature is the highest of the accurate (legitimate) temperatures that may be obtained, so entities that report this measure will have aligned motivation to estimate temperatures that are as close to the core body temperature as

possible. In one sense, the measure was designed with a compromise to pragmatism and can be thought of as having designed in a 0.5° "discount" in that our data suggest that optimal outcomes are obtained at 37.0° Celsius, rather than at the 36.5 in the measure (which is still far preferable to cooler). Further, hypothermic infants should be managed clinically using core body temperature, so there is further clinical alignment for the use of a method that approximates core body temperature.

Data from our pretesting supports various aspects of this measure. Data are from the New York State neonatal database includes reports from 20 Level 2 nurseries, 27 Level 3 nurseries, and 14 Regional Perinatal Centers that contributed 20 or more infants for the reporting year assessed. In our data we included all inborn infants from these hospitals with a birthweight of 400-2,499 grams whose admission temperature was 29° Celsius or higher (thus excluding potential data errors). Excluded were those with anencephaly or those who expired within 48 hours without receiving respiratory support beyond oxygen in the NICU. N=7,553. The number of infants ranged from 21 to 370 per hospital and 86.7% were admitted to Level 3 or higher hospitals. For this work, we used the first temperature on admission to a Level 2 or higher nursery for those admitted within 24 hours of birth.

We found that infants were <34.5 (cold), 9.6% above 34.5 but < 35.5 (very cool), 48.0% above 35.5 but < 36.5 (cool), 37.9% above 36.5 but < 37.5 (euthermic or about right), and 2.6% above 37.5 (overly warm).

There were only 67 newborns that were transferred in from another facility. The distributions of temperatures were similar to the inborn infants, with the exception the transferred infants were slightly more likely to be euthermic.

Of the inborn infants, the temperatures ranged from 29.0 to 39.7. See Table 2 below.

Table 2

Quantile	Estimate
99	37.9
95	37.3
90	37.1
75	36.8
50	36.4
25	36.0
10	35.4
5	35.0
1	34.1

The median was 36.4, the mean was 36.3, and the standard deviation was 0.7 with an interquartile range of 0.80.

Only four infants arrived in the Level 2 or higher nurseries from the Emergency Department. One infant was euthermic, one was cool, and two were very cool. Nearly 1% were transferred from the Newborn Nursery, of which 48% were euthermic, 44% cool, and only 6% very cool. None were cold.

We did not have delivery location in the dataset and therefore classified neonates born by C-section or deliveries of multiple gestations as being born in the operating room (5,254) and the remainder were classified as being born in a labor and delivery room/ birthing room (2,245). Of those born in the operating room, 2% were cold, 11% were very cool, 72% were cool, and 35% were euthermic. Those born in the L&D suite were warmer with 2% cold, 7% very cool, 13% cool, 48% euthermic, and 45% too warm ($p<.0001$). This suggests that our categorization of babies born in the OR (while imperfect) does identify a distinct population. Our Expert Panel recommended that we report by site of delivery. This is an optional stratification in our specifications.

We found that temperatures varied by birthweight category ($p<.0001$) considering those <1,000 grams, 1,000-1,499 grams, and 1,500-2,499 grams, as suggested by our Expert Panel. The percent cold was over 10% for those under 1,000 g (two-thirds of all cold babies from a group that was about 12% of all babies). These infants also were least likely to be euthermic; only 25% were so classified, compared to 34% of those in the intermediate weight category and 41% of the larger babies.

Using the categories defined in this measure, in hospital deaths were disproportionately represented among cooler babies. 2.6% of babies died before discharge: 24.5% of cold; 5.4% of very cool, 2.2% of cool babies, and 1.4% of euthermic babies died. Of the overly warm babies, 1.6% died. Only 20% of deaths came from euthermic infants.

Key findings from our study of 7,553 neonates (from 61 nurseries) in New York State are the following: temperature was variable within weight categories, and blacks were disproportionately cool compared with Hispanic or non-Hispanic others, who were disproportionately cool compared with non-Hispanic whites, whether or not we stratified by birthweight category. Deaths were disproportionate among those who were cool, in a graded fashion.

We found that performance varied by institution. Each hospital presented data for one nursery. The distribution of mean temperature by nursery ranged from 35.7 to 38.2, with a median of 36.3, a standard error of 0.36, and an interquartile range of 0.4. Twenty-five percent of these nurseries had a mean temperature below 36.1. Temperature performance varies across facilities, with less than one quarter of facilities averaging in the euthermic range.

References:

i. Dombkowski, K. J., Wasilevich, E. A., & Lyon-Callo, S. K. (2005). Pediatric asthma surveillance: Using Medicaid claims. Public Health Reports, 120, 515.

ii. Brook, R.H., et al., *A method for the detailed assessment of the appropriateness of medical technologies.* International journal of technology assessment in health care, 1986. **2**(01): p. 53-63.

2b2.3. What were the statistical results from validity testing? (*e.g., correlation; t-test*)

Our study of three hospitals found an association of hospital with lower temperatures and of lower temperatures with mortality were all significant using a generalized linear model approach at $p<.01$. The rank of the three hospitals remained constant whether we considered 1, 2, or all 3 of the lower temperature categories and was the same as the rank for mortality. Considering the proportion very cool or cold, the three hospitals scored 0.344, 0.381, and 0.414. The mortality proportions respectively were 0.051, 0.104, and 0.131, consistent with our expectations. We expect that poor

thermal management is both a direct contributor to these findings and a marker or proxy for other suboptimal aspects related to the meticulous management needed to optimize outcomes in these delicate and vulnerable infants.

2b2.4. What is your interpretation of the results in terms of demonstrating validity? (i.*e., what do the results mean and what are the norms for the test conducted?*)

Temperature is a valid and meaningful outcome. The categorical and continuous descriptions of the distribution of temperatures are each valid.

2b3. EXCLUSIONS ANALYSIS
NA ☐ no exclusions — *skip to section 2b4*

2b3.1. Describe the method of testing exclusions and what it tests (*describe the steps—do not just name a method; what was tested, e.g., whether exclusions affect overall performance scores; what statistical analysis was used*)

Exclusions include neonates with anencephaly, who receive only comfort care in the Level 2 or higher nursery, or those who die or are placed intentionally on a pre-existing hypothermia protocol prior to the 15 minute after arrival specification time.

2b3.2. What were the statistical results from testing exclusions? (*include overall number and percentage of individuals excluded, frequency distribution of exclusions across measured entities, and impact on performance measure scores*)

Because the exclusions represent a distinct population with a higher risk of short term death, such analyses would be inappropriate. In our three hospital study, we excluded 31 of 746 (4.15%).

2b3.3. What is your interpretation of the results in terms of demonstrating that exclusions are needed to prevent unfair distortion of performance results? (*i.e., the value outweighs the burden of increased data collection and analysis. Note: **If patient preference is an exclusion**, the measure must be specified so that the effect on the performance score is transparent, e.g., scores with and without exclusion*)

Exclusions are needed based upon substance and this decision is not subject to a reasonable statistical analysis, other than reporting that the number of excluded infants is small, as described above.

2b4. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES
If not an intermediate or health outcome, or PRO-PM, or resource use measure, skip to section 2b5.

2b4.1. What method of controlling for differences in case mix is used?
☐ **No risk adjustment or stratification**
☐ **Statistical risk model with** Click here to enter number of factors **risk factors**
☒ **Stratification by** Click here to enter number of categories **risk categories 3**
☐ **Other,** Click here to enter description

2b4.2. If an outcome or resource use measure is not risk adjusted or stratified, provide rationale and analyses to demonstrate that controlling for differences in patient characteristics (case mix) is not needed to achieve fair comparisons across measured entities.

The scatterplots and guidelines both justify the idea that all virtually all neonates should be able to be kept warm with appropriate management in an appropriately warm environment. The WHO Guideline describes a chain of warmth to help to guide how this may be accomplished.

For improvement and accountability purposes stratification by weight class and by race/ethnicity are specified. The IOM calls out equity as one of six fundamental components of quality. Without stratification, it is not possible to describe equity. Management of children across weight spectrums will vary normatively. In order to make the information more actionable for quality improvement, we specify stratification by weight class. Neither are intended for risk adjustment purposes. They provide additional nuance to the primary measurement, which is not adjusted.

2b4.3. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors (clinical factors or sociodemographic factors) used in the statistical risk model or for stratification by risk (*e.g., potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of p<0.10; correlation of x or higher; patient factors should be present at the start of care*)

The Pediatric Quality Measures Program that created our Center of Excellence was funded under an act of Congress that asked for measures to identify racial/ethnic and other disparities. Our race/ethnicity stratification accomplishes this.

The same CHIPRA legislation asked for reporting of performance considering urbanicity and poverty as well. So we have specified such stratifications as optional for those accountability entities that choose to use them.

The WHO guidelines recommend that all infants at all sized should be kept warm and our experts agreed. Hence we do not report only birthweight stratified results. Our expert panel felt that for reporting purposes there would be additional valuable information that to be gained by specifying weight classes for reporting such stratification. Thus we have specified such stratifications as optional for those accountability entities that choose to use them.

2b4.4a. What were the statistical results of the analyses used to select risk factors?

2b4.4b. Describe the analyses and interpretation resulting in the decision to select SDS factors (e.g. prevalence of the factor across measured entities, empirical association with the outcome, contribution of unique variation in the outcome, assessment of between-unit effects and within-unit effects)

2b4.5. Describe the method of testing/analysis used to develop and validate the adequacy of the statistical model or stratification approach (*describe the steps—do not just name a method; what statistical analysis was used*)

Provide the statistical results from testing the approach to controlling for differences in patient characteristics (case mix) below.
If stratified, skip to 2b4.9

2b4.6. Statistical Risk Model Discrimination Statistics (*e.g., c-statistic, R-squared*)**:**

2b4.7. Statistical Risk Model Calibration Statistics (*e.g., Hosmer-Lemeshow statistic*):

2b4.8. Statistical Risk Model Calibration – Risk decile plots or calibration curves:

2b4.9. Results of Risk Stratification Analysis:
Exemplar results from the statewide NY analysis:

Percent in Each Temperature Category by Race/Ethnicity Category

Percent is column percent

	White	Black	Hispanic	Other	Sum	Row Percent
Cold	1.1%	3.1%	2.1%	2.4%	149	2.0%
V Cool	4.6%	13.0%	9.7%	11.0%	644	8.6%
Cool	46.4%	50.9%	50.0%	46.9%	3621	48.5%
OK	44.5%	31.1%	36.2%	36.9%	2860	38.3%
Too Warm	3.3%	1.8%	2.1%	2.7%	193	2.6%
	N = 3051	N = 1962	N = 1793	N = 661	Total = 7467	

P < 0.01 by chi-squared analysis, "OK" is the "About Right" category

For Level 3 Nurseries, distribution of temperature by birthweight category:

Table of temprange by bwt2					
		bwt2			
		300- 999g	1000-1499g	1500-2499g	Total
temprange					
abovenorm	Frequency	9	19	59	87
	Percent	0.27	0.57	1.76	2.59
	Row Pct	10.34	21.84	67.82	
	Col Pct	1.73	2.82	2.73	
cold	Frequency	49	11	15	75
	Percent	1.46	0.33	0.45	2.23
	Row Pct	65.33	14.67	20.00	
	Col Pct	9.44	1.63	0.69	
cool	Frequency	226	336	1009	1571
	Percent	6.73	10.01	30.07	46.81
	Row Pct	14.39	21.39	64.23	
	Col Pct	43.55	49.85	46.65	
norm warm	Frequency	152	242	940	1334
	Percent	4.53	7.21	28.01	39.75
	Row Pct	11.39	18.14	70.46	
	Col Pct	29.29	35.91	43.46	
very cool	Frequency	83	66	140	289
	Percent	2.47	1.97	4.17	8.61
	Row Pct	28.72	22.84	48.44	
	Col Pct	15.99	9.79	6.47	
Total	Frequency	519	674	2163	3356
	Percent	15.46	20.08	64.45	100.00

Statistics for Table of temprange by bwt2

Statistic	DF	Value	Prob
Chi-Square	8	218.6800	<.0001
Likelihood Ratio Chi-Square	8	169.0349	<.0001
Mantel-Haenszel Chi-Square	1	0.3900	0.5323

2b4.10. What is your interpretation of the results in terms of demonstrating adequacy of controlling for differences in patient characteristics (case mix)? (i.*e., what do the results mean and what are the norms for the test conducted*)

As described above this is not relevant for our stratification. That is not its purpose. Its purpose is to report additional information.

2b4.11. Optional Additional Testing for Risk Adjustment (*not required, but would provide additional support of adequacy of risk model, e.g., testing of risk model in another data set; sensitivity analysis for missing data; other methods that were assessed*)

2b5. IDENTIFICATION OF STATISTICALLY SIGNIFICANT & MEANINGFUL DIFFERENCES IN PERFORMANCE

2b5.1. Describe the method for determining if statistically significant and clinically/practically meaningful differences in performance measure scores among the measured entities can be identified (*describe the steps—do not just name a method; what statistical analysis was used? Do not just repeat the information provided related to performance gap in 1b)*

The distribution of mean temperature by hospital ranged from 35.7 to 38.2, with a median of 36.3, a standard error of 0.36, and an interquartile range of 0.4. Twenty-five percent had a mean temperature below 36.1. We conclude that temperatures do vary across hospitals, with a range of ~7 standard errors. This is statistically meaningful variation.

2b5.2. What were the statistical results from testing the ability to identify statistically significant and/or clinically/practically meaningful differences in performance measure scores across measured entities? (e.g., *number and percentage of entities with scores that were statistically significantly different from mean or some benchmark, different from expected; how was meaningful difference defined*)

Scores are statistically varied in a manner that makes important clinical differences and that are associated with differential mortality.

2b5.3. What is your interpretation of the results in terms of demonstrating the ability to identify statistically significant and/or clinically/practically meaningful differences in performance across measured entities? (i.*e., what do the results mean in terms of statistical and meaningful differences?*)

Statistically and clinically meaningful differences are evident and captured by the measure.

2b6. COMPARABILITY OF PERFORMANCE SCORES WHEN MORE THAN ONE SET OF SPECIFICATIONS

If only one set of specifications, this section can be skipped.

Note: *This item is directed to measures that are risk-adjusted (with or without SDS factors)* ***OR*** *to measures with more than one set of specifications/instructions (e.g., one set of specifications for how to identify and compute the measure from medical record abstraction and a different set of specifications for claims or eMeasures). It does not apply to measures that use more than one source of data in one set of specifications/instructions (e.g., claims data to identify the denominator and medical record abstraction for the numerator).* ***Comparability is not required when comparing performance scores with and without SDS factors in the risk adjustment model. However, if comparability is not demonstrated for measures with more than one set of specifications/instructions, the different specifications (e.g., for medical records vs. claims) should be submitted as separate measures.***

2b6.1. Describe the method of testing conducted to compare performance scores for the same entities across the different data sources/specifications (*describe the steps—do not just name a method; what statistical analysis was used*)

2b6.2. What were the statistical results from testing comparability of performance scores for the same entities when using different data sources/specifications? (*e.g., correlation, rank order*)

2b6.3. What is your interpretation of the results in terms of the differences in performance measure scores for the same entities across the different data sources/specifications? (i.*e., what do the results mean and what are the norms for the test conducted*)

2b7. MISSING DATA ANALYSIS AND MINIMIZING BIAS

Answered in S22.

2b7.1. Describe the method of testing conducted to identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias (*describe the steps—do not just name a method; what statistical analysis was used*)

2b7.2. What is the overall frequency of missing data, the distribution of missing data across providers, and the results from testing related to missing data? (*e.g., results of sensitivity analysis of the effect of various rules for missing data/nonresponse; if no empirical sensitivity analysis, identify the approaches for handling missing data that were considered and pros and cons of each*)

2b7.3. What is your interpretation of the results in terms of demonstrating that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias**?** (i.*e., what do the results mean in terms of supporting the selected approach for missing data and what are the norms for the test conducted; if no empirical analysis, provide rationale for the selected approach for missing data*)

These are critical data for the management of newborn infants and should not be missing.

3. Feasibility

Extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

3a. Byproduct of Care Processes

For clinical measures, the required data elements are routinely generated and used during care delivery (e.g., blood pressure, lab test, diagnosis, medication order).

3a.1. **Data Elements Generated as Byproduct of Care Processes.**

Generated or collected by and used by healthcare personnel during the provision of care (e.g., blood pressure, lab value, diagnosis, depression score), Abstracted from a record by someone other than person obtaining original information (e.g., chart abstraction for quality measure or registry)

If other:

3b. Electronic Sources

The required data elements are available in electronic health records or other electronic sources. If the required data are not in electronic health records or existing electronic sources, a credible, near-term path to electronic collection is specified.

3b.1. **To what extent are the specified data elements available electronically in defined fields?** *(i.e., data elements that are needed to compute the performance measure score are in defined, computer-readable fields)*

Some data elements are in defined fields in electronic sources

3b.2. **If ALL the data elements needed to compute the performance measure score are not from electronic sources, specify a credible, near-term path to electronic capture, OR provide a rationale for using other than electronic sources.**

We have developed a web portal that is capable of collecting the data needed for this measure. It is feasible and not expensive to develop. Definition of key fields including arrival time in Nursery, and time, temperature, and location taken, as well as birth weight and time of birth (which are frequently available electronically already) would close much of the gap needed to make this an electronic measure.

3b.3. **If this is an eMeasure, provide a summary of the feasibility assessment in an attached file or make available at a measure-specific URL.**

Attachment:

3c. Data Collection Strategy

Demonstration that the data collection strategy (e.g., source, timing, frequency, sampling, patient confidentiality, costs associated with fees/licensing of proprietary measures) can be implemented (e.g., already in operational use, or testing demonstrates that it is ready to put into operational use). For eMeasures, a feasibility assessment addresses the data elements and measure logic and demonstrates the eMeasure can be implemented or feasibility concerns can be adequately addressed.

3c.1. **Describe what you have learned/modified as a result of testing and/or operational use of the measure regarding data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, time and cost of data collection, other feasibility/implementation issues.**

IF a PRO-PM, consider implications for both individuals providing PROM data (patients, service recipients, respondents) and those whose performance is being measured.

To determine the availability and ease of collecting these data elements, CAPQuaM used three primary sources: a feasibility survey of 13 hospitals conducted by The Joint Commission under contract to CAPQuaM, analysis of the Mount Sinai Data Warehouse, and a New York Statewide neonatal database that is a part of a voluntary statewide effort championed by the New York State Department of Health. The 13 hospitals included in the feasibility assessment were geographically and clinically diverse sites and were at varying stages of EMR development. The surveys were completed by the quality improvement team at each hospital. Results of these surveys revealed that the data elements required for these measures (or the information required to calculate the data element, e.g. age of neonate at time of temperature) are available at the hospital level within existing medical record systems and are not difficult to abstract.

For delivery characteristics, respondents indicated that information would be available on the infant's record, with most elements also available on the mother's record. The EMR was the preferred source of such data elements. For all other items, 12 hospitals indicated that the data were not difficult to collect, and none said that it was unavailable. A similar pattern of responses was seen regarding questions about identifying the date and time of delivery and of arrival to the intensive care nursery. Times at which the

measurement was taken (rather than the time of documentation) were universally described as present. In general, the required data elements were reported to be not difficult to collect (12/13). Data on the infant (e.g. birthweight, 5 minute Apgar score) were said to be in all of the EMRs. EMR data was seen as available to identify those managed for comfort care only and 12 hospitals indicated that such data would not be difficult to collect. Depending upon the data element, 11-13 of the sites said that race and ethnicity data and payment source would be available from the EMR. Two sites indicated that there would be a challenge to linking an infant's chart to the mother's chart, with more than 80% of the others indicating that such linkages can be performed electronically.

Analysis of the Mount Sinai Data Warehouse found that temperatures and time of temperature are often available in the Epic EMR. We found our ICD-9 schema was capable of identifying LBW infants. Some of the codes not specifically associated with a birthweight (e.g. growth retardation) were less specific for identifying LBW neonates. Details are discussed in the validation section. Of the hospitals that participate in the New York State neonatal database (using New York State designations), 23 of 25 (92.0%) classified as Level 2 nurseries submit temperature data, 31 of 36 (86.1%) with a Level 3 designation submit temperature data, and 16 of 18 (88.6%) of Regional Perinatal Centers submit temperature data. These data are virtually complete for those institutions that submit data. These data capture 84.1% of low birthweight admissions to Level 2 or higher nurseries in one year. Medicaid represents nearly half of babies entered into the database. We conclude that the necessary data are available at the level of the hospital and that such data could be collected by health plans or Medicaid programs or other entities with contractual arrangements with the providing hospitals.

The data required for the CAPQuaM perinatal measures are generally available in the existing data systems. We cannot comment on the readiness of systems to provide routine output into a database suitable for analysis and generation of these measures, but there are not fundamental barriers to such being accomplished. We have developed a stand-alone web-based data entry portal that supports this measure for all infants admitted to a Level 2 or higher nursery. We have also designed an interface that works with current EHR systems (Epic) at Mount Sinai to collect the necessary data as a part of normal workflow.

3c.2. Describe any fees, licensing, or other requirements to use any aspect of the measure as specified *(e.g., value/code set, risk model, programming code, algorithm)*.
There are no fees, licensing, or other requirements to use the measure as specified at this time.

4. Usability and Use

Extent to which potential audiences (e.g., consumers, purchasers, providers, policy makers) are using or could use performance results for both accountability and performance improvement to achieve the goal of high-quality, efficient healthcare for individuals or populations.

4a. Accountability and Transparency

Performance results are used in at least one accountability application within three years after initial endorsement and are publicly reported within six years after initial endorsement (or the data on performance results are available). If not in use at the time of initial endorsement, then a credible plan for implementation within the specified timeframes is provided.

4.1. Current and Planned Use

NQF-endorsed measures are expected to be used in at least one accountability application within 3 years and publicly reported within 6 years of initial endorsement in addition to performance improvement.

Planned	Current Use (for current use provide URL)
Not in use Use Unknown	Quality Improvement (Internal to the specific organization) Quality Improvement (Internal to the specific organization); Mount Sinai Medical Center. Implementation in progress. Ian Holzman, MD

4a.1. For each CURRENT use, checked above, provide:

- Name of program and sponsor
- Purpose
- Geographic area and number and percentage of accountable entities and patients included

4a.2. If not currently publicly reported OR used in at least one other accountability application (e.g., payment program, certification, licensing) what are the reasons? (*e.g., Do policies or actions of the developer/steward or accountable entities restrict access to performance results or impede implementation?*)

The measure is in the process of being implemented for routine quality measurement at the Mount Sinai Medical Center. We do not yet have feedback on its use.

The measure is not currently in widespread use because it is newly developed and awaiting NQF endorsement.

4a.3. If not currently publicly reported OR used in at least one other accountability application, provide a credible plan for implementation within the expected timeframes -- any accountability application within 3 years and publicly reported within 6 years of initial endorsement. (*Credible plan includes the specific program, purpose, intended audience, and timeline for implementing the measure within the specified timeframes. A plan for accountability applications addresses mechanisms for data aggregation and reporting.*)

As a part of our work with PQMP, we are working on specific plans for dissemination and use. Out plan for implementation includes submitting our application for measurement endorsement from the National Quality Forum. We are having conversations with partners regarding the application and use of this measure. No time frames have been established, as the measure will benefit from endorsement before it is implemented. Meeting the expected timeframes of NQF, the plan will include an accountability application within 3 years of initial endorsement and will be publicly reported within six years of initial endorsement.

4b. Improvement

Progress toward achieving the goal of high-quality, efficient healthcare for individuals or populations is demonstrated. If not in use for performance improvement at the time of initial endorsement, then a credible rationale describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.

4b.1. Progress on Improvement. (Not required for initial endorsement unless available.)

Performance results on this measure (current and over time) should be provided in 1b.2 and 1b.4. Discuss:

- **Progress (trends in performance results, number and percentage of people receiving high-quality healthcare)**
- **Geographic area and number and percentage of accountable entities and patients included**

n/a

4b.2. **If no improvement was demonstrated, what are the reasons? If not in use for performance improvement at the time of initial endorsement, provide a credible rationale that describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.**

This measure is clearly defined and readily understandable to consumers, patients, clinicians, providers, purchasers, health plans, policy makers and others. Throughout development, CAPQuaM brought together diverse stakeholders—clinicians, scientists, payers, purchasers, consumer organizations, and others—to ensure their iterative engagement in advancing quality measures that are understandable, salient, and actionable. CAPQuaM employed a 360° method designed to involve key stakeholders in meaningful ways.

CAPQuaM integrated perspectives from a national consortium, Steering Committee, and Senior Advisory Board at each step of the process, in addition to a continuing collaboration with AHRQ. Our team far exceeded the required minimums for expertise outside of the mainstream medical system, ensuring understandability at various levels, and by a variety of audiences.

Alpha testing was performed to assess feasibility, mechanisms of data collection, and operational aspects of collecting and analyzing data for the measure.

Our senior advisory board advised us to use lay terms in naming the measures to the extent that we are able to. This measure uses lay terminology to characterize infants who are "cold," "very cool," "cool,", "about right" (about the right temperature), and "overly warm".

The measure is specified to support both accountability and QI functions, is on a clinically critical phenomenon that is amenable to management, and is meaningful at both individual and population levels. Interest in the measure has been high.

4c. Unintended Consequences

The benefits of the performance measure in facilitating progress toward achieving high-quality, efficient healthcare for individuals or populations outweigh evidence of unintended negative consequences to individuals or populations (if such evidence exists).

4c.1. **Were any unintended negative consequences to individuals or populations identified during testing; OR has evidence of unintended negative consequences to individuals or populations been reported since implementation? If so, identify the negative unintended consequences and describe how benefits outweigh them or actions taken to mitigate them.**

There has not been any evidence of unintended negative consequences to individuals or populations. There are no anticipated unintended consequences if measuring at the level of comparing states, geographic regions, payment models, or health plans. When comparing hospitals it may be important to incorporate strata that define context, such as rurality, poverty, and types of insurance.

An exemplar presentation of the distribution aspect of the measure and an exemplar pie chart of the categorical presentation of the measure.

New CHIPRA Quality Measures on Inpatient Perinatal Quality from the Pediatric Quality Measures Program (PQMP): Hypothermia in Low Birthweight Infants

EA Howell, I Holzman, MA Kacica, EP Shields, EC Barrow, N Massenburg, CJ Homer, LC Kleinman

Mount Sinai Collaboration for Advancing Pediatric Quality Measures (CAPQuaM); Dept of Health Evidence & Policy, Icahn School of Medicine at Mount Sinai; National Institute for Child Health Quality (NICHQ), NY State Dept.of Health

PURPOSE

To describe 4 new measures developed for the federal Pediatric Quality Measures Program by the Collaboration for Advancing Pediatric Quality Measures (CAPQuaM), an AHRQ-CMS CHIPRA Center of Excellence. CHIPRA = Child Health Insurance Program Reauthorization Act (2009)

To describe the CAPQuaM approach

Consortium Partners include: AAP, AAFP, NICHQ, ACOG, CAHMI, NYS Medicaid, NCQA, Institute for Patient- and Family-Centered Care

METHODS

1. Qualitative Interviews with Clinicians
2. Literature Review
3. Criteria and Measure development
   - National multidisciplinary 9-person expert panel
   - 2 Round modified Delphi Process
   - Ratings of 81 clinical scenarios
   - Stakeholder Review and Input
4. Assessment of 4 Perinatal / hypothermia measures

Hypothermia in Premature Infants

- Associated with undesirable outcomes
  - Death
  - Intracranial hemorrhage
- Recognized in 19th Century
- Focus since Silverman, 1954
- Prevalent even in NICHD nurseries (Laptook)
- Data suggests outcomes worsen as temperature decreases from 37 degrees

Four CAPQuaM CHIPRA Perinatal Measures

Perinatal Measure 1: Percent of LBW infants who have a temperature recorded within 60 minutes of birth

Perinatal Measure 2: Percent of LWB infants who have temperature recorded within 15 minutes after arrival to NICU

Perinatal Measure 3: Distribution of temperatures for LBW neonates admitted to Level 2 or higher nurseries in the first 24 hours of life

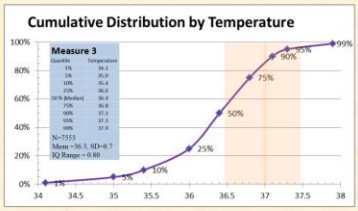
Perinatal Measure 4: Thermal condition of LBW neonates admitted to Level 2 or higher nurseries in the first 24 hours of life: above 37.5, 36.5-37.5, 35.5-36.5, 34.5-35.5, and <=34.5 (Too warm, about right, cool, very cool and cold)

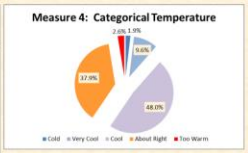
All are specified to be stratified by: birth weight category, race/ ethnicity, insurance type, urbanicity of mother's home, percent poverty in county

NY Statewide Neonatal Nursery Database analysis:

Table 1: Measures 1 and 2

Nursery Level	Temp <15 minutes After Arrival	Temp < 60 minutes after birth	Babies (N)
Level 2	87%	87%	1004
Level 3	92%	91%	3194
Level 4	82%	88%	3358
Overall	**87%**	**89%**	**7554**





FUNDING: AHRQ 1U18HS020518

Clinician Perspectives

Barriers to keeping babies warm:

Infant sick & small

System problems

- culture of late notification / failure to prep equipment & staff
- equipment failures
- lack of effective protocols
- hospital layout and L&D suite design

GUIDING PRINCIPLES

- Temperature has impact over the range of hypothermia and is not dichotomous
  - Measures using multiple lenses may be complementary
- Process of care matters

CONCLUSIONS

- CAPQuaM Perinatal hypothermia measures incorporated high levels of engagement
- Measures perform well in NY State Neonatal database: feasible and appear valid
- All measures suggest substantial opportunity for improvement and need for attention

IMPLICATIONS

- CAPQuaM measures are now available for quality measurement and QI
- Capturing the distribution of measures
  - Avoids need for falsely precise binary definition of hypothermia
  - Supports QI across range of temperatures

5. Comparison to Related or Competing Measures

If a measure meets the above criteria and there are endorsed or new related measures (either the same measure focus or the same target population) or competing measures (both the same measure focus and the same target population), the measures are compared to address harmonization and/or selection of the best measure.

5. Relation to Other NQF-endorsed Measures
Are there related measures (conceptually, either same measure focus or target population) or competing measures (conceptually both the same measure focus and same target population)? If yes, list the NQF # and title of all related and/or competing measures.
No

5.1a. **List of related or competing measures (selected from NQF-endorsed measures)**

5.1b. **If related or competing measures are not NQF endorsed please indicate measure title and steward.**

5a. Harmonization

The measure specifications are harmonized with related measures;
OR
The differences in specifications are justified

5a.1. **If this measure conceptually addresses EITHER the same measure focus OR the same target population as NQF-endorsed measure(s):**
Are the measure specifications completely harmonized?

5a.2. **If the measure specifications are not completely harmonized, identify the differences, rationale, and impact on interpretability and data collection burden.**

5b. Competing Measures

The measure is superior to competing measures (e.g., is a more valid or efficient way to measure);
OR
Multiple measures are justified.

5b.1. **If this measure conceptually addresses both the same measure focus and the same target population as NQF-endorsed measure(s):**
Describe why this measure is superior to competing measures (e.g., a more valid or efficient way to measure quality); OR provide a rationale for the additive value of endorsing an additional measure. (Provide analyses when possible.)

Appendix

A.1 Supplemental materials may be provided in an appendix. All supplemental materials (such as data collection instrument or methodology reports) should be organized in one file with a table of contents or bookmarks. If material pertains to a specific submission form number, that should be indicated. Requested information should be provided in the submission form and required attachments. There is no guarantee that supplemental materials will be reviewed.
Attachment **Attachment:** Appendix_03_21_16.docx

Contact Information

Co.1 **Measure Steward (Intellectual Property Owner):** University Hospitals Cleveland Medical Center
Co.2 **Point of Contact:** Lawrence, Kleinman, drlarrykleinman@gmail.com, 617-669-3357-
Co.3 **Measure Developer if different from Measure Steward:** Collaboration for Pediatric Quality Measures (CAPQuaM)
Co.4 **Point of Contact:** Lawrence, Kleinman, drlarrykleinman@gmail.com, 617-669-3357-

Additional Information

Ad.1 **Workgroup/Expert Panel involved in measure development**

Provide a list of sponsoring organizations and workgroup/panel members' names and organizations. Describe the members' role in measure development.

Role: Expert Panelists
Chris Lupold S - Strasburg Family Health Center
George Hoehn - Marshfield Clinic
Girija Natarajan - Children´s Hospital of Michigan Wayne State University
Jean Schied - University of Chicago Comer Children´s Hospital
Jennifer Kloesz - Magee-Womens Hospital NICU Univeristy of Pittsburgh Medical Center
Jennifer Ustianov - The University of Vermont
Jodi Jackson - Children´s Mercy Hospital University of Missouri-Kansas City School of Medicine Shawnee Mission Medical Center
Robin Knobel - Duke University School of Nursing
Thomas Bartman - The Ohio State University and Nationwide Children´s Hospital
Natalie Massenburg - Mount Sinai School of Medicine

ROLE: Steering Committee and Investigator
Wilson Pace, MD American Academy of Family Physicians – DARTNET Institute - University of Colorado
Lynn Olson, PhD American Academy of Pediatrics
Christina Bethell, PhD, MBA, MPH Child and Adolescent Health Measurement Initiative, Johns Hopkins University (Previous OHSU)
Elizabeth Howell, MD Icahn School of Medicine at Mount Sinai
Harold Kaplan, MD Icahn School of Medicine at Mount Sinai
Lawrence Kleinman, MD, MPH Icahn School of Medicine at Mount Sinai
Rebecca Anderson Mount Sinai Medical Center
Ira Nash, MD Northshore - Long Island Jewish Medical Center (previous Mount Sinai School of Medicine)
Eyal Shemesh, MD Icahn School of Medicine at Mount Sinai
Mary Barton, MD National Committee on Quality Assurance
Charles Homer, MD, MPH US Department of HHS (previous National Institute for Child Health Quality)
Marla Clayman, PhD American Institutes for Research (previous Northwestern University)
Foster Gesten, MD New York State Dept. of Health, Office of Health Insurance Programs
Jerod M. Loeb, PhD The Joint Commission
Robert Rehm National Committee for Quality Assurance
Steve Kairys, MD American Academy of Pediatrics/QuIIN
Erin DuPree, MD The Joint Commission (previous: Icahn School of Medicine at Mount Sinai)
David Baker The Joint Commission (Previous Northwestern)
Beverley Johnson, BSN* Institute for Patient- and Family-Centered Care
Doris Peter Consumers Union

ROLE: Senior Advisory Board Member and Investigator
Shoshanna Sofaer, DrPH American Institutes for Research (previous CUNY Baruch)
Lynne Richardson, MD Icahn School of Medicine at Mount Sinai
Ian Holzman, MD Icahn School of Medicine at Mount Sinai
Marilyn Kacica, MD New York State Dept. of Health, Division of Family Health

ROLE: Senior Advisory Board Member
Marc Lashley, MD Allied Pediatrics
Gary Mirkin, MD Allied Pediatrics
Barbara Kupferman, RN, CMCN, LNC AmeriChoice by United Healthcare
L. Gregory Pawlson, MD, MPH SE Healthcare QUALITY Consultants (previous Blue Cross Blue Shield)
John Santa, MD* (previous Consumers Union)
John Clarke, MD ECRI, PA Patient Safety Authority
Scott Breidbart, MD Empire Blue Cross Blue Shield/ Anthem
Jeff Terry, MBA GE Healthcare

Robert St. Peter, MD, MPH Kansas Health Institute
Ruth Stein, MD Montefiore Children's Hospital
Arthur Aufses, MD Icahn School of Medicine at Mount Sinai
Eric Rose, MD Icahn School of Medicine at Mount Sinai
Wendy Brennan, MS National Alliance on Mental Illness of New York City (NAMI - NYC Metro)
Laurel Pickering, MPH Northeast Business Group on Health
Tony Hope, BM BCh, PhD Oxford University, The Ethox Center
Martin Hatlie, JD Partnership 4 Pt. Safety, Consumers Advancing Patient Safety
Paul Wise, MD Stanford University
Mark Chassin, MD, MPH, MPP The Joint Commission
Ira Nash, MD Icahn School of Medicine at Mount Sinai
ROLE: Investigators/Key Staff
Rusty McLouth, MS AAFP
Kasey McCracken, MPH CAHMI
Elise Barrow, MPH Icahn School of Medicine at Mount Sinai
Kasey Coyne, BA Icahn School of Medicine at Mount Sinai
Natalia Egorova, PhD Icahn School of Medicine at Mount Sinai
Elizabeth Howell, MD Icahn School of Medicine at Mount Sinai
Harold Kaplan, MD Icahn School of Medicine at Mount Sinai
Melissa Manice, MPH Icahn School of Medicine at Mount Sinai
Joshua Oppenheimer, BA Icahn School of Medicine at Mount Sinai
Eve Oster, MD Pediatric Partners of Gwinnett (previous Icahn School of Medicine at Mount Sinai)
Daniel O'Sullivan, BS University of Virginia (previous Icahn School of Medicine at Mount Sinai)
Michael Parides, PhD Icahn School of Medicine at Mount Sinai
Barbara Rabin, MHA Icahn School of Medicine at Mount Sinai
Carolyn Rosen, MD Icahn School of Medicine at Mount Sinai
Melissa Saperstein, MSW Icahn School of Medicine at Mount Sinai
Eyal Shemesh, MD Icahn School of Medicine at Mount Sinai
Andrew Ting, MD Icahn School of Medicine at Mount Sinai
Adam Vella, MD Icahn School of Medicine at Mount Sinai
Virginia Walther, MSW Icahn School of Medicine at Mount Sinai
Jennifer Zeitlin, SD Icahn School of Medicine at Mount Sinai, French Institute of Health and Medical Research (INSERM)
Marianne McPherson, PhD, MS NICHQ
Joseph Anarella, MPH NYS DOH, OHIP
Lee Sanders, MD, MPH Stanford University
Kasey Coyne Northwestern
Victoria Wagner New York State Health Department Office Quality Patient Safety
Laura Soloway New York State Health Department Office Quality Patient Safety
Kevin Malloy New York State Health Department Office Quality Patient Safety
Judy Stribling Icahn School of Medicine at Mount Sinai
Linda Radecki AAP
Keri Thiessen AAP – QuIIN
Laura Morris New York State Department of Health
Olivia Lindly OHSU
Kasey McCracken OHSU
Sandeep Sharma Icahn School of Medicine at Mount Sinai
Ann Nevar UH Rainbow Babies and Children's Hospital/ CAPQuaM
Suzanne Lo UH Rainbow Babies and Children's Hospital/ CAPQuaM

Amy Balbierz Icahn School of Medicine at Mount Sinai
Samantha Raymond Icahn School of Medicine at Mount Sinai
Allisyn Vachon Icahn School of Medicine at Mount Sinai
Shannon Weber Icahn School of Medicine at Mount Sinai
ROLE: Contractors
Mary Fialkowski, MBA The Joint Commission
Linda Hanold, MHSA The Joint Commission

Elvira Ryan, MBA, BSN, RN The Joint Commission Sharon Sprenger, RHIA, CPHQ, MPA The Joint Commission Ann Watt, MBA, RHIA The Joint Commission Amy Slott The Joint Commission Denise Krusenoski The Joint Commission Elvira Ryan The Joint Commission Tasha Mearday The Joint Commission
Measure Developer/Steward Updates and Ongoing Maintenance Ad.2 **Year the measure was first released:** Ad.3 **Month and Year of most recent revision:** Ad.4 **What is your frequency for review/update of this measure?** Ad.5 **When is the next scheduled review/update for this measure?**
Ad.6 **Copyright statement:** Ad.7 **Disclaimers:**
Ad.8 **Additional Information/Comments:**

MEASURE WORKSHEET

This document summarizes the evaluation of the measure as it progresses through NQF's Consensus Development Process (CDP). The information submitted by measure developers/stewards is included after the Brief Measure Information, Preliminary Analysis, and Pre-meeting Public and Member Comments sections.

To navigate the links in the worksheet: Ctrl + click link to go to the link; ALT + LEFT ARROW to return

Brief Measure Information

NQF #: 2896
Measure Title: Structural Attributes of Facility in which High Risk Women Deliver Newborns: A PQMP Measure
Measure Steward: University Hospitals Cleveland Medical Center
Brief Description of Measure: This measure characterizes the facility that is the site of delivery of newborn infants born to high risk women by four key structural characteristics. These four characteristics were identified as critical structures by a national expert panel who served CAPQuaM's 360 degree process for measure development. This work was undertaken in the context of developing innovative measures of the availability of High Risk Obstetrical (HROB) care as assigned by AHRQ and CMS.

The four key structures are:

(a) Level 3 or higher NICU services on campus. Level 3 NICU is defined as meeting either the American Academy of Pediatrics (AAP) criteria or a locally used set of explicit criteria recognized by that state's Department of Health.

(b) 24/7 on-site blood banking services/transfusion services that are always available for obstetrical patients. By 24/7 blood banking/transfusion services we mean that the following are always available to obstetrical patients: testing of blood group and Rh Type; cross matching; antibody testing; transfusion with on-site and available blood, either ABO specified or O-Rh-negative; transfusion with fresh frozen plasma; and transfusion with cryoprecipitate.

(c) 24/7 in - house physician dedicated to labor and delivery who is capable of safely managing labor and delivery, and of performing a cesarean section, including an emergent cesarean section.

(d) 24/7 in - house physician coverage dedicated to the obstetrical service by an anesthesiologist who is qualified to provide obstetrical anesthesia.
Developer Rationale: Each of the components has independent information that may inform improvement. As an accountability measures, this is an all or none composite.

Numerator Statement: Number of eligible newborn deliveries that occur in facilities with:
(a) Level 3 or higher NICU services on campus. Level 3 NICU is defined as meeting either the American Academy of Pediatrics (AAP) criteria or a locally used set of explicit criteria recognized by that state's Department of Health.
(b) 24/7 on-site blood banking services/transfusion services that are always available for obstetrical patients. By 24/7 blood banking/transfusion services we mean that the following are always available to obstetrical patients: testing of blood group and Rh Type; cross matching; antibody testing; transfusion with on-site and available blood, either ABO specified or O-Rh-negative; transfusion with fresh frozen plasma; and transfusion with cryoprecipitate.
(c) 24/7 in - house physician dedicated to labor and delivery who is capable of safely managing labor and delivery, and of performing a cesarean section, including an emergent cesarean section.
(d) 24/7 in - house physician coverage dedicated to the obstetrical service by an anesthesiologist who is qualified to provide obstetrical anesthesia.

Measure: Meets all four criteria.

Stratifications:
a. Meets none
b. Includes a
c. Includes b
d. includes c
e. includes d

Numerator Elements:
Number of eligible deliveries
Maternal and infant ICD-9 codes
Response to survey question identified on technical specifications or Other valid self-report of structural characteristics as specified

No Numerator Exclusions
Denominator Statement: Overall number of newborn deliveries in health care facilities that are born to women whose pregnancy meets the criteria for high risk. While qualification for the denominator requires that the birth occur in a health care facility this measure is not specified to assess performance of individual facilities.
Denominator Exclusions: None

Measure Type: Composite
Data Source: Administrative claims, Healthcare Provider Survey
Level of Analysis: Health Plan, Integrated Delivery System, Population : Community, Population : County or City, Population : National, Population : Regional, Population : State

IF Endorsement Maintenance – Original Endorsement Date: Most Recent Endorsement Date:

New Measure -- Preliminary Analysis

Criteria 1: Importance to Measure and Report

1a. Evidence

1a. Evidence. The evidence requirements for a structure, *process or intermediate outcome* measure is that it is based on a systematic review (SR) and grading of the body of empirical evidence where the specific focus of the evidence matches what is being measured.

The developer provides the following evidence for this measure:

- **Systematic Review of the evidence specific to this measure** ☐ **Yes** ☒ **No**
- **Quality, Quantity and Consistency of evidence provided?** ☐ **Yes** ☒ **No**
- **Evidence graded?** ☐ **Yes** ☒ **No**

Evidence Summary

- The developer provides a conceptual model diagram of the relationships of the structures to outcomes.
- The developer cites a 2015 article that identified "Domains of childbirth services potentially related to childbirth outcomes include 24 hours availability of obstetrician/family practitioner physician, anesthesia staffing (e.g. practioner type, structure, 24 hour availability), NICU capacity 24 hours and licensed level of ICU care, and Blood bank availability 24 hours." (Korst, 2015). Korst concludes that linkage of this data to childbirth outcomes should assist in the identification of key resources and activities that optimize hospital environment.
- This composite measure has four components. Evidence for each components is described:
  - **Level 3 or higher NICU:** Multiple guidelines and selected studies cited. American Academy of Pediatrics Committee on Fetus and Newborn. ACOG Committee on Obstetric Practice. Riley AE, Stark AR, Kilpatrick SJ, Papile LA, eds. Guidelines for Perinatal Care, 7th Ed. Elk Grove Village, IL, AAP 2012. Page 78:"All attempts to be made to ensure that women and infants at high risk receive care at a facility that

provides the required level of obstetric and newborn care. " This is the only component where empirical studies were cited.

- **Blood banking/transfusion 24/7:** Guidelines from ACOG, AAP and National Partnership for Maternal Safety all stress the need to for availability of blood products at all times.
- **In-house laborist/C-section 24/7**: *Levels of Obstetric Care. Obstetrics Care Consensus No. 2*. American College of Obstetricians and Gynecologists. Obstet Gynecol 2015:125:502-515 Level 2 or higher delivery systems (1 is lowest, 4 highest) should have ob-gyn at all times.
- **Dedicated OB anesthesia 24/7:** American Academy of Pediatrics Committee on Fetus and Newborn. ACOG Committee on Obstetric Practice. Riley AE, Stark AR, Kilpatrick SJ, Papile LA, eds. Guidelines for Perinatal Care, 7th Ed. Elk Grove Village, IL, AAP 2012. Page 26:" A Board certified anesthesiologist with special training or experience in maternal fetal anesthesia should be in charge of obstetric anesthesia services...' at hospitals caring for high risk pregnancies. Obstetrical anesthesia "should be available in the hospital 24 hours a day".

- The Appendix contains a summary of a targeted literature review :

  This literature review examined the evidence for the use of the following availability measures in obstetrics: 24-hour obstetrical availability, 24-hour anesthesia availability, access to maternal fetal medicine specialist, access to subspecialty care, and access to multidisciplinary care. It was determined that when 24 hour obstetrical services are not available, there may be an increase in morbidity and mortality as the incidence of intra partum anoxia may increase. Twenty-four hour availability of anesthesia for high risk obstetrical deliveries has caused a great deal of discussion among the medical community. Concerns lie in the current availability of anesthesia services in the United States, the workforce estimate of staffing for the future, and the barriers to providing immediate availability in all hospitals that provide obstetrical care.

- The developer "used a Scoping Review of the literature and coupled it with a RAND type expert panel using a modified Delphi approach" to create the measure construct.
- The evidence is primarily ungraded guidelines and selected studies.

Exception to evidence

Because the evidence for this measure is mostly expert consensus rather than a systematic review of empirical evidence, it is insufficient to meet NQF's criterion for evidence. However, an exception to the evidence criterion is allowed if the Committee agrees that empirical evidence is not needed to hold providers accountable for the measure.

Guidance from the Evidence Algorithm

Process measure (Box 1) --> not based on a SR (Box 3) --> no empirical evidence (Box 7) --> systematic assessment of expert opinion (Box 11) --> if Committee agrees it is OK/beneficial to hold providers accountable for performance in the absence of empirical evidence of benefits to patients → rate as INSUFFICIENT WITH EXCEPTION

Staff NOTE: The Committee will first vote on the criterion. If a>60% votes for insufficient then a second vote will be taken to determine whether the Committee wishes to pass the evidence criterion with exception.

Questions for the Committee:

- *For possible exception to the evidence criterion:*
  - *Are there, or could there be, performance measures of a related health outcome, OR evidence-based intermediate clinical outcomes, intervention/treatment?*
  - *Is there evidence of a systematic assessment of expert opinion beyond those involved in developing the measure?*
  - *Does the SC agree that it is acceptable (or beneficial) to hold providers accountable without empirical evidence?*

Preliminary rating for evidence: ☐ **High** ☐ **Moderate** ☐ **Low** ☒ **Insufficient**

1b. Gap in Care/Opportunity for Improvement and 1b. Disparities
Maintenance measures – increased emphasis on gap and variation

1b. Performance Gap. The performance gap requirements include demonstrating quality problems and opportunity for improvement.

- On page 6 of the Appendix the developers provide two graphs of % of high risk deliveries in hospitals with Level 3 OB care and % of high risk deliveries in hospitals with NICU using MAX Medicaid data and AHA categorization of hospitals for 18 states. The developers note that the "AHA definitions are not as strict as the specifications on our measure."
- Results are mostly below 50% for all states for Level 3 OB care.
- Results are mostly below 60% for NICU.

Disparities

- During testing of the measure the developers found several areas of disparities.
- The developers found that "In Large Metropolitan areas among women who met our criteria for high-risk deliveries, 44.76% of black women, 40.11% of Hispanic women, and 30.04% of white women in Medicaid delivered in hospitals with Level 3 or higher NICUs." in New York state.
- The developers found a different pattern with regional perinatal centers (proxy for 24/7 blood banking/transfusion centers) where for large metro areas among women who met other criteria for high-risk deliveries, 19.25% of white women, 13.92% of black women, and 13.82% of Hispanic women deliver at these institutions.
- The developers examined high risk deliveries by poverty levels using the upper three quartiles based on USDA definitions in New York State:

HROB Summary (Combined Unduplicated) New York State Medicaid, 2010

Poverty Level	N	OB Proxy*	Transfusion Proxy **	NICU >=3***
Top Quartile	8,533	20.27%	12.83%	25.08%
Second Quartile	44,013	26.58%	14.42%	37.89%
Third Quartile	3,919	10.56%	2.91%	9.95%

- Additional analyses were performed for rural vs. urban.

Questions for the Committee:

- *Is there a gap in care that warrants a national performance measure?*
- *Can this measure provide useful data on disparities in perinatal care?*

Preliminary rating for opportunity for improvement: ☒ High ☐ Moderate ☐ Low ☐ Insufficient

1c. Composite - Quality Construct and Rationale

1c. Composite Quality Construct and Rationale. The quality construct and rationale should be explicitly articulated and logical; a description of how the aggregation and weighting of the components is consistent with the quality construct and rationale also should be explicitly articulated and logical.

- This is an all-or-none composite measure of four structural components.
- The developer cites a 2015 article that identified "Domains of childbirth services potentially related to childbirth outcomes include 24 hours availability of obstetrician/family practitioner physician, anesthesia staffing (e.g. practitioner type, structure, 24 hour availability), NICU capacity 24 hours and licensed level of ICU care, and blood bank availability 24 hours." (Korst, 2015)
- The developer concludes that "Normatively, these composites should move together with Centers that deliver high risk infants having all of these. In our assessment we did not find that they moved in unison and there is additive information. Patient safety is best supported by having each of the services on site at all times for deliveries that are above normal risk."

Questions for the Committee:

- *Are the quality construct and a rationale for the composite explicitly stated and logical?*
- *Is the method for aggregation and weighting of the components explicitly stated and logical?*

Preliminary rating for composite quality construct and rationale:

☒ **High** ☐ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 1: Importance to Measure and Report (including 1a, 1b, 1c)

1a. Evidence to Support Measure Focus

Comments:

Composite measure with insufficient evidence to support.

1b. Performance Gap

Comments:

Yes, there are gaps in care. Need a broader perspective to determine the exact disparities but different patterns are noted. Data from California and New York.

1c. Composite Performance Measure

Comments:

All or nothing measure, and the overall quality construct is logical. Component performance measures are stated. Appears section 1a.7 was not completed as required. No aggregation or weighting rules.

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability

2a1. Reliability Specifications

Maintenance measures – no change in emphasis – specifications should be evaluated the same as with new measures

2a1. Specifications requires the measure, as specified, to produce consistent (reliable) and credible (valid) results about the quality of care when implemented.

Data source(s): Administrative claims, Healthcare Provider Survey (included in specifications)

Specifications:

- Level of analysis is health plan, integrated delivery system and population: community, region, state
- The numerator is number of eligible newborn deliveries that occur in facilities with all four components (maintain a separate count of those who meet each of the four criteria for stratification purposes.)
- The denominator is the number of newborn deliveries in health care facilities that are born to women whose pregnancy meets the criteria for high risk as defined in two ways:
  - Class A: Maternal Diagnoses and Comorbidities
  - Class B: Delivery Complications, Fetal Injury or Compromise, or Suboptimal Infant Diagnoses
    - Maternal Delivery Complication Codes (ICD9)
    - Maternal Stillbirth or Birth Hypoxia/Asphyxia Codes
    - Premature or small infant (Infant codes)
  - The Unduplicated union of Class A and Class B is the denominator for this measure.
  - ICD 9 and ICD 10 codes for Class A and B are included in the attached data dictionary spreadsheet.
- The measure is stratified for each of the components as a separate result.
- Survey questions for hospital characteristics is specified.
- A calculation algorithm is provided.

Questions for the Committee:

- *Are all the data elements clearly defined? Are all appropriate codes included?*
- *Is the logic or calculation algorithm clear?*
- *Is it appropriate to include delivery complications or birth asphyxia as a marker for high-risk for the denominator?*
- *Is it likely this measure can be consistently implemented?*

2a2. Reliability Testing Testing attachment

2a2. Reliability testing demonstrates if the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise enough to distinguish differences in performance across providers.

Reliability testing level ☐ **Measure score** ☒ **Data element** ☐ **Both**
Reliability testing performed with the data source and level of analysis indicated for this measure ☒ **Yes** ☐ **No**

Method(s) of reliability testing

- The developer states "The data collection and reliability therein depend upon the use of administrative data. These data are used to identify deliveries (our specifications are a slight enhancement of CDC methodologies described in Kuklina et. al; to Kuklina's work we added Revenue code 722). The CDC is using this method as its standard for identifying deliveries from administrative data." The developers chose to "specify the measure based upon the Kuklina/CDC approach as it is both widely-used and relevant for the type of population-based approach to measurement proposed in this measure."
- The developers note that "In determining which women were to be considered potentially in need of HROB services, our specifications further rely upon administrative data."
- The developer reports "Regarding the assessment of the presence or absence of structural characteristics in this measure set, we have specified this measure to use the results of questionnaires or surveys that we envision as paper, email or internet-based. Our feasibility assessment determined that these data are readily available from key individuals at the hospitals."

Results of reliability testing

- The developer cites four studies of administrative data element validity when compared to the medical record. None are specific to the codes in this measure.
- The developers conclude that "Since this measure is specified to be interpreted at the population and not the individual level, the impact of some of the imperfections of using administrative data will be overcome naturally because of the law of large numbers."
- No empirical testing of the specific data elements used in this measure is presented.

Guidance from the Reliability Algorithm

Precise specifications (Box 1) → empirical reliability testing done – no (Box 2) → empirical validity testing of data elements – no (Box 3) → insufficient

Questions for the Committee:

- *Does this measure meet the criterion for empirical reliability testing?*
- *Has the reliability of the hospital survey been established?*
- *Is the test sample adequate to generalize for widespread implementation?*
- *Do the results demonstrate sufficient reliability so that differences in performance can be identified?*

Preliminary rating for reliability: ☐ **High** ☐ **Moderate** ☐ **Low** ☒ **Insufficient**

2b. Validity

2b1. Validity: Specifications

2b1. Validity Specifications. This section should determine if the measure specifications are consistent with the evidence.

Specifications consistent with evidence in 1a. ☒ **Yes** ☐ **Somewhat** ☐ **No**

Question for the Committee:

- *Are the specifications consistent with the evidence?*

2b2. Validity testing

2b2. Validity Testing should demonstrate the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality.

SUMMARY OF TESTING

Validity testing level ☐ **Measure score** ☒ **Data element testing against a gold standard** ☐ **Both**

Method of validity testing of the measure score:

- ☐ **Face validity only**
- ☐ **Empirical validity testing of the measure score**

Validity testing method:

- The developers identified a hypothesis:
  - ***Hypothesis:*** All else equal, we would expect women who live in more medically dense communities to experience greater availability than those who live in less medically dense communities and those who live in more isolated communities to have even less availability.
- Developer studied the relationship between anesthesia services and hospital characteristics.
- Developer reports on study the relationship between of delays in obtaining blood products and hospital size from California.

Validity testing results:

  - In data from New York State Medicaid among women who met the criteria for high-risk deliveries, the developers found that these measures vary with a gradient of accessibility of medical services associated with geographic proximity or metropolitan areas. See Table 6.
  - Variability in anesthesia staffing is not defined by hospital characteristics.
  - Hospital size did not correlate with availability of blood products.
- Developer conclusion: "We interpret the findings to suggest that these services become less available with increasing rurality, as we had predicted. We designed the measures to identify reduced availability for any reason (including geographic isolation) and the observed gradient strongly supports the validity of these as population measures of availability. These components move in similar directions but not in lock step, confirming that they are measuring related but not identical constructs,"
- None of the data were analyzed at the level of plans or health systems.

Questions for the Committee:

- *Is the test sample adequate to generalize for widespread implementation?*
- *Is validity sufficiently established to be used in plans and health systems as well as states?*
- *Do the results demonstrate sufficient validity so that conclusions about quality can be made?*
- *Do you agree that the score from this measure as specified is an indicator of quality?*

2b3-2b7. Threats to Validity

2b3. Exclusions:

- The specifications indicate there are no denominator exclusions.

- The developer notes "We compared the number of eligible births in NY State with Class A and Class B to those with Class A only and conclude there is no need to exclude if maternal and infant charts cannot be linked."

2b4. Risk adjustment: **Risk-adjustment method** ☒ **None** ☐ **Statistical model** ☐ **Stratification**

- The developer notes "We call for stratification by race/ethnicity as called for in the CHIPRA legislation that funded our work. We further specify for optional stratification by Poverty and Rurality that we recommend but do not require. These are not for purposes of risk adjustment but allow for more granular understanding of the data and the detection of racial and ethnic disparities."

2b5. Meaningful difference *(can statistically significant and clinically/practically meaningful differences in performance measure scores can be identified)*:

- The developer reports that "Analysis of variance using generalized models and SAS software demonstrated state to state differences, racial and ethnic differences, and differences between rural, urban, and suburban areas for the constructs that make up this measure."
- State data is present in graphs for two components of the composite. No data is presented using the full composite measure results for either states or plans/systems.

Question for the Committee:

- *Does this measure identify meaningful differences about quality?*
- *Is there enough information to determine whether this measure will be useful for plans or health systems?*

2b6. Comparability of data sources/methods: NA

2b7. Missing Data

- In S.22. the developer indicates that missing data are considered "no" responses.

2d. Composite measure: construction

2d. Empirical analysis to support composite construction. Empirical analysis should demonstrate that the component measures add value to the composite and that the aggregation and weighting rules are consistent with the quality construct.

- For an all-or-none composite measures the analysis should include performance rates of the various components.
- The developer did not answer these questions. The developer has addressed some of these issues throughout the submission for some of the components.

Questions for the Committee:

- *Do the component measures fit the quality construct?*
- *Are the objectives of parsimony and simplicity achieved while supporting the quality construct?*

Guidance from the algorithm:
Specifications consistent with the evidence (Box 1) → potential threats to validity addressed – no measure results; no composite measure analysis → insufficient

Preliminary rating for validity: ☐ **High** ☐ **Moderate** ☐ **Low** ☒ **Insufficient**

Committee pre-evaluation comments

Criteria 2: Scientific Acceptability of Measure Properties (including all 2a, 2b, and 2d)

2a1. & 2b1. Specifications

Comments:

Data elements are clearly defined. Codes are included. Delivery complications may arise from normal pregnancies. If these occur, the providers and/or facilities should not be held accountable for this patient not delivering in a high-risk facility. Unsure if this measure would be consistently implemented.

Data element is tested against a gold standard.

2a2. Reliability Testing

Comments:

Did not have empirical reliability testing. Insufficient reliability.

2b2. Validity Testing

Comments:

The testing needs to be conducted outside of New York. Don't think the results have enough validity so that conclusions about quality nationally can be made.

2b3. Exclusions Analysis

2b4. Risk Adjustment/Stratification for Outcome or Resource Use Measures

2b5. Identification of Statistically Significant & Meaningful Differences In Performance

2b6. Comparability of Performance Scores When More Than One Set of Specifications

2b7. Missing Data Analysis and Minimizing Bias

Comments:

Yes. If data is missing, it is reported as "no". This could skew the results.

2d. Composite Performance Measure

No. No aggregation or weighting in this measure.

Criterion 3. Feasibility

Maintenance measures – no change in emphasis – implementation issues may be more prominent

3. Feasibility is the extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

- The developer states "The measure was developed in tandem with feasibility assessment and is structured to be feasible and readily captured. We learned that in lieu of the survey measures, it was possible to use proxies which were better for some (NICU) than others (not available for anesthesiologist). A regular survey and registry of these features would make measurement easier in an ongoing fashion rather than periodic collection of the survey as would be necessary today."
- Administrative data is generally thought to be feasible.

Questions for the Committee:

- *Are the required data elements routinely generated and used during care delivery?*
- *Is it reasonable to expect plans/systems and states/communities to collect the survey data from every hospital serving their population?*
- *Are the required data elements available in electronic form, e.g., EHR or other electronic sources?*
- *Is the data collection strategy ready to be put into operational use?*

Preliminary rating for feasibility: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments
Criteria 3: Feasibility

3a. Byproduct of Care Processes
3b. Electronic Sources
3c. Data Collection Strategy
Comments:
** Each facility's structure and available resources need to be determined, then data could be reported out.**

Criterion 4: Usability and Use

4. Usability and Use evaluate the extent to which audiences (e.g., consumers, purchasers, providers, policymakers) use or could use performance results for both accountability and performance improvement activities.

Current uses of the measure
Publicly reported? ☐ **Yes** ☒ **No**

Current use in an accountability program? ☐ **Yes** ☒ **No**
OR
Planned use in an accountability program? ☐ **Yes** ☐ **No**

Accountability program details
- The developers are "making plans for dissemination and use" but not details are provided.

Improvement results new measure

Unexpected findings (positive or negative) during implementation NA

Potential harms

Feedback:

***Questions for the Committee**:*
- *How can the performance results be used to further the goal of high-quality, efficient healthcare?*
- *Do the benefits of the measure outweigh any potential unintended consequences?*

Preliminary rating for usability and use: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments
Criteria 4: Usability and Use

4a. Accountability and Transparency
4b. Improvement
4c. Unintended Consequences
Comments:
**Not publicly reported. Nationally this would help further define gaps in care and needed resources.
Don't know of any unintended consequences.**

Criterion 5: Related and Competing Measures
Related or competing measures none **Harmonization**

Pre-meeting public and member comments

-

NATIONAL QUALITY FORUM—Evidence (subcriterion 1a)

Measure Number (*if previously endorsed*)**:** Click here to enter NQF number
Measure Title: Structural Attributes of Facility in which High Risk Women Deliver Newborns: A PQMP Measure
IF the measure is a component in a composite performance measure, provide the title of the Composite Measure here: Click here to enter composite measure #/ title

Date of Submission: 3/22/2016

Instructions

- *For composite performance measures:*
  - *A separate evidence form is required for each component measure unless several components were studied together.*
  - *If a component measure is submitted as an individual performance measure, attach the evidence form to the individual measure submission.*
- Respond to all questions as instructed with answers immediately following the question. All information needed to demonstrate meeting the evidence subcriterion (1a) must be in this form. An appendix of *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 10 pages (*incudes questions/instructions*; minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.

Note: The information provided in this form is intended to aid the Steering Committee and other stakeholders in understanding to what degree the evidence for this measure meets NQF's evaluation criteria.

1a. Evidence to Support the Measure Focus
The measure focus is evidence-based, demonstrated as follows:

- Health outcome: [3] a rationale supports the relationship of the health outcome to processes or structures of care. Applies to patient-reported outcomes (PRO), including health-related quality of life/functional status, symptom/symptom burden, experience with care, health-related behavior.
- Intermediate clinical outcome: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured intermediate clinical outcome leads to a desired health outcome.
- Process: [5] a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured process leads to a desired health outcome.
- Structure: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured structure leads to a desired health outcome.
- Efficiency: [6] evidence not required for the resource use component.

Notes

3. Generally, rare event outcomes do not provide adequate information for improvement or discrimination; however, serious reportable events that are compared to zero are appropriate outcomes for public reporting and quality improvement.
4. The preferred systems for grading the evidence are the U.S. Preventive Services Task Force (USPSTF) grading definitions and methods, or Grading of Recommendations, Assessment, Development and Evaluation (GRADE) guidelines.
5. Clinical care processes typically include multiple steps: assess → identify problem/potential problem → choose/plan intervention (with patient input) → provide intervention → evaluate impact on health status. If the measure focus is one step in such a multistep process, the step with the strongest evidence for the link to the desired outcome should be selected as the focus of measurement. Note: A measure focused only on collecting PROM data is not a PRO-PM.
6. Measures of efficiency combine the concepts of resource use and quality (see NQF's Measurement Framework: Evaluating Efficiency Across Episodes of Care; AQA Principles of Efficiency Measures).

1a.1.This is a measure of: (*should be consistent with type of measure entered in De.1*)
Outcome

☐ Health outcome: Click here to name the health outcome
☐Patient-reported outcome (PRO): Click here to name the PRO

PROs include HRQoL/functional status, symptom/symptom burden, experience with care, health-related behaviors

☐ Intermediate clinical outcome (*e.g., lab value*): Click here to name the intermediate outcome

☐ Process: Click here to name the process

☒ Structure: (a) with level 3 or higher NICU services on campus
(b) with 24/7 in-house blood banking/transfusion services available
(c) 24/7 in-house physician capable of safely managing labor and delivery, and performing a cesarean section, including an emergent cesarean section, and
(d) with 24/7 in-house physician coverage dedicated to the obstetrical service by an anesthesiologist who is qualified to provide obstetrical anesthesia,

☐ Other: Click here to name what is being measured

HEALTH OUTCOME/PRO PERFORMANCE MEASURE *If not a health outcome or PRO, skip to 1a.3*

1a.2. Briefly state or diagram the path between the health outcome (or PRO) and the healthcare structures, processes, interventions, or services that influence it.

1a.2.1. State the rationale supporting the relationship between the health outcome (or PRO) to at least one healthcare structure, process, intervention, or service (*i.e., influence on outcome/PRO*).

Note: For health outcome/PRO performance measures, no further information is required; however, you may provide evidence for any of the structures, processes, interventions, or service identified above.

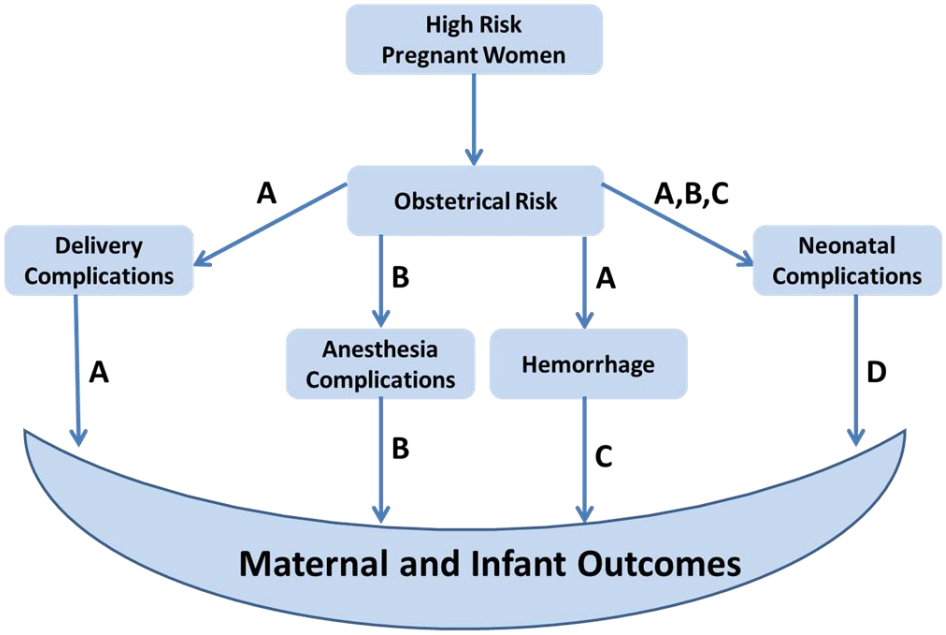
INTERMEDIATE OUTCOME, PROCESS, OR STRUCTURE PERFORMANCE MEASURE

1a.3. Briefly state or diagram the path between structure, process, intermediate outcome, and health outcomes. Include all the steps between the measure focus and the health outcome.

The current measure addresses four critical structures necessary for optimal outcomes among women with high-risk conditions. While there is much interest in obstetrics in classifying levels of obstetric care, we prioritized four specific attributes that others might use to define such levels (2).

High Risk deliveries may be complicated by obstetrical emergencies, requiring skillful management of labor and delivery including urgent c-section, various complex needs for anesthesia benefit from an obstetrical anesthesiologist, hemorrhage requires instantly available blood banking services, and critically ill infants, requiring NICU services, often including an expert in neonatal resuscitation in the delivery room.. The use of such structural components as we propose is consistent with the emerging thinking in women's health care and at national societies and federal agencies regarding levels of care.

Please see our conceptual model for causality below:



This measure considers 4 structural attributes of delivering facilities that may enhance maternal and fetal outcomes by preventing or managing complications.
A. 24/7 coverage of Labor and Delivery services by a physician skilled in obstetrical care and capable of performing an emergency C-section may prevent or manage delivery complications, avert hemorrhage, and enhance neonatal outcomes. **B.** 24/7 coverage of the L&D service by an anesthesiologist can prevent or manage anesthetic complications and improve both maternal and infant outcomes. **C.** 24/7 blood banking services allows for rapid transfusion during and after hemorrhage and improves clinical outcomes for women and infants. **D.** On campus neonatal intensive care improves neonatal outcomes and reduces the need for potentially dangerous transport.

Examining many components of childbirth services, Korst et al. examined professional standards, including the Guidelines for Perinatal Care, published by the American Academy of Pediatrics and the American College of Obstetricians and Gynecologists, medical literature and worked with perinatologists and perinatal nurse managers to assure that the breadth of each domain studied was well represented. Domains of childbirth services potentially related to childbirth outcomes include 24 hours availability of obstetrician/family practitioner physician, anesthesia staffing (e.g. practioner type, structure, 24 hour availability), NICU capacity 24 hours and licensed level of ICU care, and Blood bank availability 24 hours. (Korst, LM, Feldman, DS., Bolman, DL., et al. Variation in childbirth services in California: a cross-sectional survey of childbirth hospitals. Am J Obstet Gynecol 2015;213:523.e1-8.)

Key references for each of these structural attributes includes:

A. A highly qualified laborist reduces complications and enhances maternal and infant outcomes. Emergency C-sections can improve outcomes.
Guideline: *Levels of Obstetric Care. Obstetetrics Care Consensus No. 2*. American College of Obstetricians and Gynecologists. Obstet Gynecol 2015:125:502-515 Level 2 or higher delivery systems (1 is lowest, 4 highest) should have ob-gyn at all times.

American Academy of Pediatrics Committee on Fetus and Newborn. ACOG Committee on Obstetric Practice. Riley AE, Stark AR, Kilpatrick SJ, Papile LA, eds. Guidelines for Perinatal Care, 7th Ed. Elk Grove Village, IL, AAP 2012

B. A dedicated obstetrical anesthesiologist can improve labor outcomes.
Guideline: *Levels of Obstetric Care. Obstetrics Care Consensus No. 2*. American College of Obstetricians and Gynecologists. Obstet Gynecol 2015:125:502-515

American Academy of Pediatrics Committee on Fetus and Newborn. ACOG Committee on Obstetric Practice. Riley AE, Stark AR, Kilpatrick SJ, Papile LA, eds. Guidelines for Perinatal Care, 7th Ed. Elk Grove Village, IL, AAP 2012

American Academy of Pediatrics Committee on Fetus and Newborn. ACOG Committee on Obstetric Practice. Riley AE, Stark AR, Kilpatrick SJ, Papile LA, eds. Guidelines for Perinatal Care, 7th Ed. Elk Grove Village, IL, AAP 2012. Page 26:" A Board certified anesthesiologist with special training or experience in maternal fetal anesthesia should be in charge of obstetric anesthesia services...' at hospitals caring for high risk pregnancies. Obstetrical anesthesia "should be available in the hospital 24 hours a day".

C. Immediate blood banking services can improve labor outcomes.
Guideline: *Levels of Obstetric Care. Obstetetrics Care Consensus No. 2*. American College of Obstetricians and Gynecologists. Obstet Gynecol 2015:125:502-515 "capabilities should [include] … available support services including access to obstetric … blood bank supplies at all times"

Guideline: D'Alton, ME., Main, EK., Menard MK., and Levy, BS. (2014). The National Partnership for Maternal Safety. Obstetrics & Gynecology. Vol. 123 (5): 973-977. Calls for "rapid and sustained availability of blood products"

[Review] Clark SL. Strategies for reducing maternal mortality. Semin Perinatol 2012;36:42–7. "insufficient or slow replacement of blood and clotting components in the patient with massive ongoing hemorrhage" represents a common failure in the management of hemorrhage that contributes to preventable maternal deaths.

Clark SL, Belfort MA, Dildy GA, et al. Maternal Death in the 21st Century: Causes, prevention, an relationship to Caesarian delivery. American J Obstetr Gynec. 2008. 199:e1-e5 found that 8 of 17 potentially preventable maternal deaths were due to failures to manage hemorrhage appropriately.

American Academy of Pediatrics Committee on Fetus and Newborn. ACOG Committee on Obstetric Practice. Riley AE, Stark AR, Kilpatrick SJ, Papile LA, eds. Guidelines for Perinatal Care, 7th Ed. Elk Grove Village, IL, AAP 2012 Page 255 discusses importance of transfusion of r hemorrhage in mother or newborn. Page 212 discusses importance of blood banking services for specific indications of high risk deliveries. Page 10 indicates that even basic delivery care requires 24/7 blood bank services.

D. NICU improves neonatal outcomes. On site NICU is preferable in terms of outcomes.

American Academy of Pediatrics Committee on Fetus and Newborn. ACOG Committee on Obstetric Practice. Riley AE, Stark AR, Kilpatrick SJ, Papile LA, eds. Guidelines for Perinatal Care, 7th Ed. Elk Grove Village, IL, AAP 2012. Page 78:"All attempts to be made to ensure that women and infants at high risk receive care at a facility that provides the required level of obstetric and newborn care. If a women is receiving obstetric care with a low level of NICU care and she is found to be at risk of adverse outcome or premature delivery, transfer of care to a hospital of high level of NICU care is indicated, wherever safely possible. Delivery hospitals that do not have a level 3 or level 4 NICU should develop affiliations…Formal transfer agreements should be in place."

Empirical data on NICU in hospital of delivery:
In an analysis of extremely low-birthweight infants, Kaneko showed that mortality and long-term outcomes are affected by the process of maternal or infant transport (Kaneko et al). A population-based study of 195 infants from 189 mothers following 50,632 deliveries looked at processes and indications for maternal or neonatal transport, maternal and infant characteristics, and the prognosis for extremely low-birthweight infants. Using multiple statistical methods, they determined that the rates of mortality and handicaps among the infants in the maternal

transport group were 15.2% and 23.2%, respectively, compared to 25% and 44% in the neonatal transport group. They concluded that the rate of morbidity and mortality of extremely low birthweight infants may be reduced by the reduction of neonatal transports.

Kaneko, M., Yamashita, R., Kai, K., Yamada, N., Sameshima, H., and Ikenoue, T. (2015) Perinatal morbidity and mortality for extremely low-birthweight infants: A population-based study of regionalized maternal and neonatal transport. J Obstet Gynaecol Res, 41: 1056–1066. doi: 10.1111/jog.12686.

A prospective study by Modanlou matched 131 neonates transported antenatally with 131 transported postnatally (comparable birth weights and gestational ages). The neonates in the antenatal group had a significantly lower incidence of respiratory distress syndrome and other morbidity and a shorter hospitalization, suggesting that antenatal transport of high-risk pregnancies to a regional perinatal center may significantly affect outcomes.

Modanlou, H D (12/1980). "Perinatal transport to a regional perinatal center in a metropolitan area: Maternal versus neonatal transport.". American journal of obstetrics and gynecology(0002-9378), 138 (8), p. 1157.
PMID: 7446624

In 2014, Jensen and Lorch performed a retrospective population-based cohort study of all VLBW infants without severe congenital anomalies delivered in all hospitals in 3 states from January 1999 through December 2009 (N = 72,431). Risk-adjusted odds ratios and risk-adjusted probabilities were determined by logistic regression. They found that risk of death or severe intraventricular hemorrhage and death or necrotizing enterocolitis was lowest among infants born in hospitals that had both a high volume of VLBW infant deliveries and a high-level NICU. Complications were also more common among infants born at hospitals with a level I or II NICU compared with infants delivered at hospitals with a level IIIB/C NICU. They demonstrated that antenatal transfer of high-risk pregnancies to hospitals with higher levels of NICU may reduce mortality and improve outcomes.

Effects of a Birth Hospital's Neonatal Intensive Care Unit Level and Annual Volume of Very Low-Birth-Weight Infant Deliveries on Morbidity and Mortality Online Only
Erik A. Jensen, MD; Scott A. Lorch, MD, MSCE
JAMA Pediatr. 2015;169(8):e151906. doi:10.1001/jamapediatrics.2015.1906.

An earlier (1996), very similar study reached the same conclusions. Concentration of high-risk deliveries in urban areas in a smaller number of hospitals that could provide level III NICU care has the potential to decrease neonatal mortality (without increasing costs).

The Effects of Patient Volume and Level of Care at the Hospital of Birth on Neonatal Mortality
Ciaran S. Phibbs, PhD; Janet M. Bronstein, PhD; Eric Buxton; Roderic H. Phibbs, MD
JAMA. 1996;276(13):1054-1059. doi:10.1001/jama.1996.03540130052029.

Noting that this study and others involved smaller samples or narrowly defined networks, and most were based on data collected before the routine use of surfactant-replacement therapy, an update was performed studying VLBW infants. Using logistic regression to estimate odds ratios for mortality associated with the NICU level of care and annual volume of VLBW infants, this study showed that NICU volume and level is strongly associated with mortality. Again, mortality was lowest for deliveries in hospitals with high-level and high-volume NICUs, though less than a quarter of VLBW infants are born in such hospitals and the percentage has been decreasing. The authors suggest that increased regionalization of perinatal care might reduce mortality.

Phibbs CS, Baker LC, Caughey AB, Danielsen B, Schmitt SK, Phibbs RH. Level and volume of neonatal intensive care and mortality in very-low-birth-weight infants. NEngl J Med 2007; 356(21): 2165–2175.

Looking at the issue of centralized (regionalized) versus decentralized perinatal care, a 2010 study concluded that despite studies showing that centralized perinatal services result in better outcomes, decentralization is occurring in some regions due to the belief that it leads to improved access to care in underserved areas. In California, the decentralization of services has resulted in an increase in the number of midlevel units and a decrease in at-risk patients delivering in units with the availability neonatal intensive care services. In this study, nearly 20% of VLBW deliveries continue to be delivered in level-1 and -2 hospitals.

Chung JH, Phibbs CS, Boscardin WJ, Kominski GF, Ortega AN, Needleman J. The effect of neonatal intensive care level and hospital volume on mortality of very low birth weight infants. Med Care 2010; 48: 635–644.

The proposed availability measures address important gaps in quality and safety and also have the potential to narrow disparities in maternal and neonatal outcomes. These four structural attributes (24-hour in-house physicians covering obstetrics and capable of managing labor and delivery, including performing emergent cesarean sections, 24-hour in house physicians available and capable of providing obstetric anesthesia, 24-hour availability of blood bank/transfusion services, and delivery at a facility with a Level 3 or higher NICU) have the potential to improve both maternal and infant outcomes in the setting of high-risk deliveries. They were chosen to represent a prioritized selection of key structural attributes that impact the timeliness with which a potentially urgent service may be available to women who are delivering in the context of a pregnancy that manifests higher than typical risk. The prioritization process involved our team of stakeholders as well as an Expert Panel, whose clinical and health services judgments guided the process.

Delivery care provided to pregnant women is critical for the health and well-being of mothers and babies. The burden of chronic illness and risk factors for pregnancy complications (e.g. hypertension, diabetes, advancing maternal age, previous cesarean section) are all rising among women, increasing their risk for morbidity and mortality (2). Over the past decade, maternal mortality has increased in the U.S.; striking racial disparities persist (3, 4). Black women are 3 to 4 times more likely to suffer a pregnancy-related death than white women (3). Racial and ethnic disparities are also reflected when considering both the processes and outcomes of neonates
(5-7).

For every maternal death, 100 or more women suffer severe maternal morbidity, a potentially life- threatening diagnosis or life-saving procedure that is associated with pregnancy. Examples include organ failure (e.g. acute renal failure, liver, respiratory), obstetric shock, pulmonary embolism, amniotic embolism, eclampsia, septicemia, cardiac events, mechanical ventilation, transfusion, invasive hemodynamic monitoring, and hysterectomy. Severe maternal morbidity is rising and affects approximately 52,000 women annually in the US (4). Studies using the Nationwide Inpatient Sample show the prevalence of at least one severe complication rose 75% from 1998-99 to 2008-09: renal failure increased by 97%, thrombolic embolism by 100%, adult respiratory distress syndrome by 75%, blood transfusion by 183%, and ventilation by 34%. Similar to maternal and neonatal mortality, minority women are more likely to suffer a severe maternal morbidity than white women (4). Severe morbidity is more common at the extremes of reproductive age and for black women as compared with white women. Quality and safety of care are an important lever to address these issues as research suggests that at least one-third to one- half of maternal deaths in the US may be preventable through improvements in quality of care (8-10). Additional studies suggest that on the continuum of care to adverse pregnancy outcomes, there are a number of points that can be impacted by improved safety and quality (11). All four structural measures are critical to ensure safety of mothers and babies in the setting of deliveries that are of higher risk, whether due to maternal comorbidities or complications of pregnancy (12,13). We refer to these collectively as high-risk deliveries. Please see 1.c.4 for citation list.

To improve care for women who require high-risk obstetrical services, it is imperative that quality measures address the availability of high-risk obstetrical services by assessing how available key services are at hospitals providing obstetric care. Agencies such as the March of Dimes, American Academy of Pediatrics, American College of Obstetricians and Gynecologists, American Academy of Family Physicians and American Medical Association have emphasized the need for stratification of facilities based on maternal levels of care, so that the definition of levels of care should be based on the capability to provide more complex care. For example, Table 1 displays the Indiana Perinatal Network's criteria on when to consult, refer, or transport a pregnant woman. See Table 1 below. Similarly, quality measures can play a critical role in identifying gaps in care delivery and subsequently act to decrease severe maternal morbidity and mortality (2).

Table 1. Indiana Perinatal Network Levels of Inpatient Obstetric Care

Level I (Basic)	Level II (Specialty)	Level III (Subspecialty)
▸ Uncomplicated labor/delivery (> 36 weeks), antepartum/intrapartum/postpartum ▸ C-section capability available 24 hours per day within 30 minutes	▸ Level-I-plus care of selected high-risk mothers and fetuses ▸ Portable ultrasound in-house and available for diagnostic visualization of fetus as well as capabilities to perform biophysical tests and amniotic fluid analysis	▸ Level-II-plus comprehensive perinatal services (management of severe maternal complications) ▸ Maternal-fetal medicine specialist on staff and available for consultation 24-hours per day
▸ Stabilization of mother for transfer	▸ Co-director of perinatal services is board-certified (qualified) obstetrician	▸ Attending OB available in-house on 24-hour basis
▸ Director or co-director of perinatal services is board-certified (qualified) obstetrician or family practice physician trained in obstetrics	▸ Director of OB anesthesia is board-certified (qualified) anesthesiologist experienced in OB anesthesia	▸ Full complement of specialists readily available (includes but not limited to surgery, infectious disease, hematology, respiratory therapy, internal medicine) 24-hours per day ▸ Genetics counselor in-house or available by referral ▸ Co-director of perinatal services is board-certified (qualified) in maternal fetal medicine ▸ Director of OB anesthesia is board-certified (qualified) anesthesiologist experienced in OB anesthesia ▸ OB anesthesia available on 24-hour basis

Reprinted from Indiana Perinatal Network. Levels of Hospital Perinatal Care in Indiana: October 2008. Available at: www.indianaperinatal.org/downloads/Levels-of-Hospital-Perinatal-Care-in-Indiana.pdf. Retrieved July 11, 2012 *and* data from the American Academy of Pediatrics and the American College of Obstetricians and Gynecologists. Guidelines for perinatal care. 6th ed. Elk Grove Village (IL) and Washington (DC); 2007.

The CAPQuaM measure development process sought to ground availability measures in a definitional framework of what constitutes a high-risk obstetrical service. First, we approached the literature to establish a construct of conditions that potentially can be considered as high risk, increasing the risk of maternal and/or infant morbidity and mortality. We subsequently convened a multidisciplinary panel of national experts to provide leadership, including helping to establish definitions for both availability and high-risk obstetrical services. The panel held a telephone meeting, conducted pre-work via email and participated in a two-day face to face meeting. By the conclusion of the meeting the Panel had rated a variety of constructs using this adaptation of the two round RAND/UCLA modified Delphi process. These measures are derived from literature and consensus and anticipated the new ACOG guideline, which makes strong recommendations (Evidence Level 1C) for meaningful impact for these structural attributes enhancing outcomes. (*Levels of Obstetric Care. Obstetetrics Care Consensus No. 2*. American College of Obstetricians and Gynecologists. Obstet Gynecol 2015:125:502-515)

This is the first measure we are aware of that addresses specifically the availability of high-risk obstetrical care. It reflects our perspective that the optimal health of children in the United States is fostered by healthy pregnancies and deliveries. The availability of HROB services is critical for the health of pregnant women with high-risk deliveries and ultimately for the health of the child they are carrying. An emerging consensus in the literature relates the construct of levels of care for women and newborn services. The American Academy of Pediatrics (AAP) defines special and intensive care newborn services as Levels 2-4 in a specific manner and the field of obstetrics is rapidly moving in that direction. These measures both build off of the AAP definition and operationalize components that comprise levels of high-risk obstetrical services. They capture the extent to which women in need of HROB services and who may be at risk for or experiencing a complicated delivery are delivered at hospitals that provide sufficient care.

Consortium partners at the New York State Department of Health, including the Office of Health Insurance Programs / New York State Medicaid, steering committee, and scientific team have played central roles to the development of these measures. Evidence for high level of interest in this work in particular was demonstrated by the fact that the CAPQuaM team was asked to present this work in development to CMS Expert Panel on Improving Maternal and Infant Health Outcomes in Medicaid/CHIP Data, Measurement, and Reporting Workgroup.

More generally, childbirth is the largest category for hospital admissions for commercial payers and Medicaid programs and the estimated annual hospital costs associated with childbirth and newborn care are over $80 billion in the United

States annually (14, 15). In New York State, 48.6% of deliveries in 2011 occurred in women insured by Medicaid (16). In our analysis year, 55.6% (4197 neonates) of low birthweight neonates admitted to NICUs across New York State and who were in our study of newborn temperatures (approximately 90% of all newborns admitted to level 2 or 3 nurseries) were insured by Medicaid.

Providing high quality care to women with high-risk deliveries has the potential both to improve outcomes and to narrow disparities, important national priorities for CMS. In fact, leaders in obstetrics have proposed systematic changes in the delivery of obstetric care to address these issues. Both peer-reviewed and grey literature propose improved integrated maternal-fetal- neonatal care networks that optimize regionalization of care to improve access to critical 24/7 in- house obstetric services, blood bank/transfusion services, obstetrical anesthesia, and level 3 or 4 NICU services for women with high-risk pregnancies (2, 12, 13).

Therefore, the proposed measure has the potential to have a significant impact on the health of mothers and infants insured by Medicaid. High-risk deliveries disproportionately impact women insured by Medicaid as compared with private insurance. Risk factors identified to be associated with high-risk deliveries (e.g., hypertension, delivery of low birth weight infants) are all factors that are more prevalent among the Medicaid population. Given the fact that childbirth is the leading category for hospital admissions for Medicaid programs and the fact that high-risk deliveries disproportionately occur among women insured by Medicaid, quality measures targeting high-risk deliveries have the potential to improve quality of care for a sizeable portion of the Medicaid program.

The Expert Panel offered definitions regarding which conditions established that a pregnancy required high-risk obstetrical services. They specifically endorsed the importance of certain services being available 24/7 in the hospital of delivery, among those a qualified obstetrical physician, an obstetrical anesthesiologist, blood banking/transfusion services, and a Level 3 or higher NICU. The conclusion long and in depth discussion by the panel led to consensus that the role of these availability measures should be to describe availability at a population level even though the unit of analysis that we were to measure directly was an individual pregnancy. There are two key implications: these measures are not intended to assess the quality of care for a given pregnancy and they are intended to generate a gradient along which availability of HROB services can be assessed. They do provide a measure of the performance of a health plan or system to the specified population of high risk women. The full nuance of this measure's capacity to describe availability will be enhanced over time by the establishment of benchmarks in medically and geographically diverse populations and communities.

The co-leads of this measure development, a pediatrician and an obstetrician, collaboratively operationalized these constructs into the measures in the current measure set, working with the CAPQuaM stakeholders—including NY Medicaid—and consulting the Expert Panelists as appropriate. Using ICD9 codes and a publicly available grouping system, AHRQ's Clinical Classification Software (http://www.hcup-us.ahrq.gov/toolssoftware/ccs/ccs.jsp), the various conditions that could classify a pregnancy as in need of HROB services were specified into those seen in this measure. The four measures in this set incorporate these high priority conditions and services and address the capacity to have immediately available high-risk services before, during and after delivery. They describe the proportion of high-risk deliveries that take place in facilities that meet one or more of four structural criteria.

1) 24/7 in-house physician staffing the obstetrical unit who is capable of safely managing labor and delivery, and performing a cesarean section, including an emergent cesarean section.
2) 24/7 in-house obstetrical anesthesia services
3) 24/7 in-house blood banking/transfusion services
4) Level 3 or higher NICU services

The New York State Office of Health Insurance Programs is an active CAPQuaM partner and has been engaged in the conceptualization and development of these measures.

Strong recommendations amidst a limited evidence base call for the presence of Ob anesthesiologists, 24-7 physicians capable of emergent surgical management of delivery, and the presence of blood banking services. Strong empirical

evidence as well as guidelines and recommendations support the observation that transport of maternal women is safer than of sick or small neonates, and that neonatal intensive care improves outcomes including mortality.

1a.3.1. What is the source of the systematic review of the body of evidence that supports the performance measure?

☒ Clinical Practice Guideline recommendation – ***complete sections*** *1a.4* ***, and*** *1a.7*

☐ US Preventive Services Task Force Recommendation – ***complete sections*** *1a.5* ***and*** *1a.7*

☐ Other systematic review and grading of the body of evidence (*e.g., Cochrane Collaboration, AHRQ Evidence Practice Center*) – ***complete sections*** *1a.6* ***and*** *1a.7*

☒ Other – ***complete section*** *1a.8*

Please complete the sections indicated above for the source of evidence. You may skip the sections that do not apply.

_____

1a.4. CLINICAL PRACTICE GUIDELINE RECOMMENDATION

1a.4.1. Guideline citation (*including date*) and **URL for guideline** (*if available online*):

Levels of Obstetric Care. Obstetetrics Care Consensus No. 2. American College of Obstetricians and Gynecologists. Obstet Gynecol 2015:125:502-515 "capabilities should [include] … available support services including access to obstetric … blood bank supplies at all times" Level 2 or higher delivery systems (1 is lowest, 4 highest) should have ob-gyn at all times. Board certified anesthesiologist with training in obstetric anesthesia available. (Table 1)

D'Alton, ME., Main, EK., Menard MK., and Levy, BS. (2014). The National Partnership for Maternal Safety. Obstetrics & Gynecology. Vol. 123 (5): 973-977. Calls for "rapid and sustained availability of blood products" in all birthing facilities.

American Academy of Pediatrics Committee on Fetus and Newborn. ACOG Committee on Obstetric Practice. Riley AE, Stark AR, Kilpatrick SJ, Papile LA, eds. Guidelines for Perinatal Care, 7th Ed. Elk Grove Village, IL, AAP 2012

1a.4.2. Identify guideline recommendation number and/or page number and **quote verbatim, the specific guideline recommendation**.

Please see American Academy of Pediatrics Guidelines for Perinatal Care under 1A.3.

1a.4.3. Grade assigned to the quoted recommendation with definition of the grade:

ACOG recommendations 1C

1a.4.4. Provide all other grades and associated definitions for recommendations in the grading system. (*Note: If separate grades for the strength of the evidence, report them in section 1a.7.*)

1a.4.5. Citation and URL for methodology for grading recommendations (*if different from 1a.4.1*)**:**

1a.4.6. If guideline is evidence-based (rather than expert opinion), are the details of the quantity, quality, and consistency of the body of evidence available (e.g., evidence tables)?

☐ Yes → ***complete section*** *1a.7*

☒ No → *report on another systematic review of the evidence in sections 1a.6 and 1a.7; if another review does not exist,* ***provide what is known from the guideline review of evidence in*** *1a.7*

1a.5. UNITED STATES PREVENTIVE SERVICES TASK FORCE RECOMMENDATION

1a.5.1. Recommendation citation (*including date*) and **URL for recommendation** (*if available online*):

1a.5.2. Identify recommendation number and/or page number and **quote verbatim, the specific recommendation**.

1a.5.3. Grade assigned to the quoted recommendation with definition of the grade:

1a.5.4. Provide all other grades and associated definitions for recommendations in the grading system. (*Note: the grading system for the evidence should be reported in section 1a.7.*)

1a.5.5. Citation and URL for methodology for grading recommendations (*if different from 1a.5.1*)**:**

Complete section 1a.7

1a.6. OTHER SYSTEMATIC REVIEW OF THE BODY OF EVIDENCE

1a.6.1. Citation (*including date*) and **URL** (*if available online*):

See appendix for our scoping review which informed this work. Additional supplemental literature were performed but are not included in the appendix.

1a.6.2. Citation and URL for methodology for evidence review and grading (*if different from 1a.6.1*)**:**

Complete section 1a.7

1a.7. FINDINGS FROM SYSTEMATIC REVIEW OF BODY OF THE EVIDENCE SUPPORTING THE MEASURE

If more than one systematic review of the evidence is identified above, you may choose to summarize the one (or more) for which the best information is available to provide a summary of the quantity, quality, and consistency of the body of evidence. Be sure to identify which review is the basis of the responses in this section and if more than one, provide a separate response for each review.

1a.7.1. What was the specific structure, treatment, intervention, service, or intermediate outcome addressed in the evidence review?

1a.7.2. Grade assigned for the quality of the quoted evidence with definition of the grade:

1a.7.3. Provide all other grades and associated definitions for strength of the evidence in the grading system.

1a.7.4. What is the time period covered by the body of evidence? (*provide the date range, e.g., 1990-2010*). Date range: Click here to enter date range

QUANTITY AND QUALITY OF BODY OF EVIDENCE

1a.7.5. How many and what type of study designs are included in the body of evidence? (*e.g., 3 randomized controlled trials and 1 observational study*)

1a.7.6. What is the overall quality of evidence across studies in the body of evidence? (*discuss the certainty or confidence in the estimates of effect particularly in relation to study factors such as design flaws, imprecision due to small numbers, indirectness of studies to the measure focus or target population*)

ESTIMATES OF BENEFIT AND CONSISTENCY ACROSS STUDIES IN BODY OF EVIDENCE

1a.7.7. What are the estimates of benefit—magnitude and direction of effect on outcome(s) across studies in the body of evidence? (*e.g., ranges of percentages or odds ratios for improvement/ decline across studies, results of meta-analysis, and statistical significance*)

1a.7.8. What harms were studied and how do they affect the net benefit (benefits over harms)?

UPDATE TO THE SYSTEMATIC REVIEW(S) OF THE BODY OF EVIDENCE

1a.7.9. If new studies have been conducted since the systematic review of the body of evidence, provide for each new study: 1) citation, 2) description, 3) results, 4) impact on conclusions of systematic review.

_____

1a.8 OTHER SOURCE OF EVIDENCE

If source of evidence is NOT from a clinical practice guideline, USPSTF, or systematic review, please describe the evidence on which you are basing the performance measure.

1a.8.1 What process was used to identify the evidence?

We used a Scoping Review of the literature and coupled it with a RAND type expert panel using a modified Delphi approach, all of this embedded in a highly engaged and transparent process that was validated via a peer review and funding process.

The CAPQuaM team conducted an initial literature review considering availability of high risk obstetrical care. Upon reviewing the findings across the CAPQuaM Consortium and working with a national expert panel that was convened for a RAND style modified Delphi process, the team moved to a more focused search to inform the development of quality measures. Three research questions were identified to address evidence-based, effective interventions related to the provision of services to high-risk obstetrical patients. They are as follows: 1) Identify the interventions, services, policies and practices that are effective in reducing morbidity and mortality in mothers and infants in 6 specific high-risk conditions (diabetes/obesity, cardiac disease, preeclampsia/eclampsia, hypertension, psychiatric/mental health disorders, and hemorrhage); 2) Elicit the findings and recommendations from review of the literature related to the 6 high-risk conditions that have been used as evidence-based guidelines or quality measures; and 3) Define specific obstetric services – availability of 24 hour obstetric and anesthesia services, and access to maternal/fetal medicine specialists, subspecialty care and multidisciplinary care - that have been addressed in the professional literature, and the impact of varying availability and accessibility on fetal and maternal morbidity and mortality.

The team then conducted an extensive search of the literature following a study protocol. Inclusion criteria were as follows: literature documenting interventions, services procedures or practices found to be effective or ineffective in reducing maternal or infant morbidity and mortality in pregnant mothers of all ages with at least one of the above 6 conditions. Exclusion criteria were as follows: literature documenting conditions not related to one of the six (above), publications dates prior to 1980, references that were duplicated, in utero co-morbidity, population not related to maternal-fetal, and a research focus not related to above three research questions. Over a two month period, the team identified a total of 3,663 literature abstracts and titles that met search criteria, analyzed 515 of these articles and included 310 in the literature review. The findings were incorporated into CAPQuaM's 360 degree process as described elsewhere. This full literature review can be found in the appendix.

Some selected findings are presented herein:

An American Hospital Association Chart Book (17) describes OB services as important to be always available; delay can impair maternal and neonatal outcomes. The failure to respond urgently and definitively to fetal distress, maternal hemorrhage or any number of complications during the peurperium can lead to sub-optimal outcomes or death. The Indiana Perinatal Network (2) considers 24/7 in house obstetrical services to be part of subspecialty care; our partners in

NY State require a maternal-fetal medicine specialist and a neonatologist always be on-site and available within 20 minutes in order to be designated either a Level 3 Perinatal Center or a Regional Perinatal Center (i.e., Level 4).

Qualifying under the OB construct measure requires coverage of OB by a physician capable of providing the indicated service. Physicians may be obstetricians or family physicians qualified to fill those roles. In testing the measure, we used a proxy which should be sensitive with moderate specificity, hospitals' self-report of being a Level 3 hospital for Obstetrics on the AHA Survey, supplemented by a NY hospital profiling website (18). In 2010 in NY State Medicaid, 24.52% of Class A deliveries, 27.98% of Class B, and 24.66% of unduplicated combined A and B occurred in hospitals that met the structural measure for OB care.

In all measures Class A and B are reported separately to promote understandability to complement the combined finding (which always will be dominated in terms of numbers by Class A).

The integration of OB anesthesia into high-risk care has become accepted (19). With increasing complexity of available anesthesia techniques (20) and an increase in the risk of deliveries occurring, our Expert Panel chose to operationalize the structural need as a 24/7 anesthesiologist with training in obstetric anesthesia. This is consistent with literature and reflects the diverse roles of the obstetrical anesthesiologist: managing pain, administering anesthesia, managing severe hypertension, and intubating and managing the complications thereof (19, 21, 22). As these measures are intended to assess availability, panelists were not dissuaded by data (20) suggesting a shortage of OB anesthesiologists. Such a shortage may motivate the use and elevate the importance of this measure. As noted above, ACOG, and AAP call for obstetrical anesthesiologist availability in their clinical guidelines and when defining levels of care for higher risk women.

The use of general anesthesia in 15-30% of emergent C-sections contrasted with less than 5% in elective C-sections simultaneously suggests that urgent situations may require different management than do routine ones (arguing for highly skilled specialists), and potentially that a lack of higher level anesthesia care in urgent situations may limit available options for women (19, 20). Interviews with clinicians during the CAPQuaM process suggest that lack of availability frequently does limits women's options.

Of interest, hospitals that deliver between 100 and 500 babies represent 36% of all hospitals and account for almost 8% of births, suggesting the importance of workforce distribution (20). They make up 36% of hospitals, suggesting their critical importance when developing measures of availability. 20% of hospitals that delivered 500-1500 deliveries per year (the middle stratum) reported themselves to be regional referral centers for HROB.

Recent data updates these findings. CAPQuaM is working with Drs. Jill Mhyre (an Expert Panel member), Andrea Fuller, and Brenda Bucklin on a manuscript, "Anesthesia Services for High-risk Obstetrics: Results from the 2011 Obstetric Anesthesia Workforce Survey." This survey supports the salience of this measure. Results are shown in Tables 2-4 below. The data within categories are nationally representative, but final sampling weights are not ready to make national estimates across categories.

TABLE 2

		High-risk referral center	Not a high-risk referral center	P-value
Facility Type				
	Community hospital	61(58.7%)	207 (91.6%)	<0.001
	Anesthesia residency	42 (40.4%)	6 (2.7%)	
	CRNA training program	1 (1%)	4 (1.8%)	
	Military	0	2 (0.9%)	
	Rural critical access	0	7 (3.1%)	
Annual Delivery volume*		3000 [2000, 4500]	750 [350,1800]	<0.001
Cesarean Deliver y rate *		30% [27, 35]	29% [25, 33]	0.01

Proportion funded by private insurance*	45% [25, 65]	35% [20, 70]	0.22

TABLE 3

	High-risk referral center	Not a high-risk referral center	P-value
Night-time and weekend staffing			
In-house anesthesiologist or anesthesia team dedicated to L&D	61(60.4%)	57 (26.3%)	<0.001
In-house anesthesiologist or anesthesia team with additional duties	25 (24.8%)	27 (12.4%)	
In-house independently-practicing CRNA dedicated to L&D	6 (5.9%)	5 (2.3%)	
In-house independently-practicing CRNA with additional duties	0	6 (2.8%)	
Anesthesiologist or CRNA on call from home	9 (8.9%)	12 2 (56.2%)	
Antepartum consultation service	89 (89.0%)	140 (65.4%)	<0.001
Massive transfusion protocol	86 (95.6%)	113 (56.0%)	<0.001

	High-risk referral center	Not a high-risk referral center	P-value
Night-time and weekend staffing			
In-house anesthesiologist or anesthesia team dedicated to L&D	61(60.4%)	57 (26.3%)	<0.001
In-house anesthesiologist or anesthesia team with additional duties	25 (24.8%)	27 (12.4%)	
In-house independently-practicing CRNA dedicated to L&D	6 (5.9%)	5 (2.3%)	
In-house independently-practicing CRNA with additional duties	0	6 (2.8%)	
Anesthesiologist or CRNA on call from home	9 (8.9%)	12 2 (56.2%)	
Antepartum consultation service	89 (89.0%)	140 (65.4%)	<0.001
Massive transfusion protocol	86 (95.6%)	113 (56.0%)	<0.001

TABLE 4

	Self-reported characteristics of the respondent's institution				
	High-risk referral center ≥1500 births per year	High-risk referral center <1500 births per year	Low risk, ≥1500 births per year	Low risk, <1500 births per year	*P-value*
Recommended anesthesia staffing					
In-house anesthesiologist	55 (68.8%)	3 (30%)	26 (44.8%)	19 (13.7%)	<0.001
In-house medically directed resident or CRNA	19 (23.8%)	2 (20%)	16 (27.6%)	15 (10.8%)	
In-house CRNA without medical direction	4 (5%)	4 (40%)	5 (8.6%)	20 (14.4%)	
In-hospital coverage unnecessary	2 (2.5%)	1 (10%)	11 (19.0%)	85 (61.2%)	
Recommended obstetric staffing					
In-house obstetrician	75 (93.8%)	8 (80%)	41 (69.5%)	43 (30.9%)	0.001
In-house nurse midwife	0	0	1 (1.6%)	5 (3.6%)	
In-house other physician	0	1 (10%)	0	2 (1.4%)	
In-hospital coverage unnecessary	5 (6.3%)	1 (10%)	17 (28.8%)	89 (64.0%)	

Variability in anesthesiology staffing is not defined by hospital characteristics. While similar structural characteristics predict obstetrical and anesthesiology coverage they do not overlap, supporting distinct measures for OB and anesthesiology coverage.

Our transfusion measure incorporates language from the NY State DOH criteria to identify Regional Perinatal Centers in NY. The clinical imperative to look at availability of these services is set forth by the California Maternal Quality Care Coalition (CMQCC) (13). Hemorrhages occur predictably in the context of coagulation disorder, somewhat predictably when problems of placentation may be noted before or early in labor, or unpredictably. Large amounts of blood loss may go unnoticed or unappreciated if not monitored, sought, and understood by experienced and meticulous clinicians, often aided by thoughtful protocols. And even in the hands of excellent clinicians, the management of hemorrhage requires early recognition, proper management to achieve rapid hemostasis, and prompt and sometimes repeated transfusion. Key data from CMQCC are shown in Table 5 below.

TABLE 5

Frequency of Delays in obtaining blood products when needed By Individual Respondents n (% by row)

Hospital Size: # Live Births (2005)	No Delay	No Blood Bank on-site	Lack of Pre-natal Record	Blood Bank hesitancy to release O-	Blood bank closed/off hours	Blood Bank is busy	Total
<1000 (33)	24 (73)	2 (6)	0 (0)	2 (6)	0 (0)	3 (9)	31 (94)
1001-3000 (121)	59 (49)	6 (5)	9 (7)	15 (12)	0 (0)	13 (11)	102 (84)
>3000 (86)	47 (55)	2 (2)	5 (6)	13 (15)	0 (0)	12 (14)	79 (92)
Total (240)	130 (54)	10 (4)	14 (6)	30 (12)	0 (0)	28 (12)	212* (88)

For our NY State Medicaid data analysis we used regional perinatal centers (RPC) as a (highly specific, moderately sensitive) proxy for round-the-clock transfusion services--RPC are required to have them always available. Among HROB deliveries, for Class A, 13.38%; Class B, 12.62%; and the combined 13.46% delivered in RPC hospitals. We note here another "voltage drop" between OB coverage and blood bank services, validating our decision to include both measures.

Regionalization of perinatal care has been widely accepted in the US; studies document that delivery at hospitals with Level III NICUs is associated with reduced neonatal mortality and the American Academy of Pediatrics encourages regionalization of NICU services (12, 23, 24) and established Level 3 NICUs as standard of care for many infants. Our 2010 New York State Medicaid analysis found that the following proportion of deliveries in hospitals that had Level 3 or higher nurseries (identified in this data set by regular submission of Revenue Code 173 or 174): Class A, 34.01%; Class B, 37.25%; and unduplicated combined 34.16%. Even for Class B, in which the desirability for a NICU is highest and most proximal, nearly 2/3 of women deliver in hospitals that do not have one.

Our literature review, data collection, and data analyses reveal many deliveries in institutions that lack desirable structural characteristics, plus the independent importance of each of these related measure. Some outcomes evidence combines with strong national recommendations to support this measure

Evidence is discussed throughout this form. A 200 page targeted review of the literature is in the Appendix. Further, we interviewed clinicians, engaged clinical societies and accreditors, patient/family groups, NY Medicaid and others to inform our measure development with the intelligence and experiences of stakeholders as well as the medical literature. The ratings of the panel along with a brief description of methodology are included as Appendices. These measures result from careful conduction of a systematic process.

The availability of high-risk obstetric (HROB) services is a challenging concept, and to develop quality measures that assess availability of high-risk obstetrics services we first needed to define: 1) availability of services and 2) high-risk obstetrical services. Specifically we wondered whether the target population could be identified by conditions present in the women, by the clinical services required, or by the clinicians providing the services. Through discussions with our Scientific Team, Steering Committee, review of the literature, and in consultation with our Expert Panel, we answered these questions in the following manner. Regarding availability, we expanded on the Anderson and Aday model (25), which suggests that utilization of health care is driven by predisposing characteristics, enabling resources and need, and that these factors are themselves influenced by the available system of care (26, 27). While their distinction between availability and realized access has blurred over time, we nonetheless chose to respect our assignment by using an availability lens as our framework for these measure.

At a system level, utilization can vary as a result of differences in individual behaviors or system characteristics. The current measures predominantly reflect distribution of system attributes, which may include geography, system design, and/or sufficiency of resources (27). The definition of HROB services for the purposes of these measures is broad and may include services provided by a variety of clinicians if received by a woman who has an identifiable condition that predisposed her or her baby to an increased risk of morbidity and mortality during the assessment period. For this

measure set, we developed two subcategories of high-risk identification. Derived from the literature, Expert Panel ratings, and discussions with our Steering Committee, and from insights drawn from clinician interviews, we include a group of maternal diagnosis codes that place women at increased risk of maternal morbidity and mortality and a group of codes that represent complications of delivery, including low birthweight, that place infants at risk for increased morbidity and mortality. Poor birth outcomes, such as birth asphyxia or stillbirth are included among the latter.

A significant proportion of pregnant women are at higher risk for maternal or infant morbidity and mortality (2). Professional societies in pediatrics, anesthesia, and obstetrics provide guidance about the need for availability of specific services regarding HROB. A 2009 Joint Statement from the American Society of Anesthesiologists (ASA) and the American College of Obstetricians and Gynecologists (ACOG) called for available OB anesthesia services. Optimal anesthesia care should include credentialed clinicians always available to administer an appropriate anesthetic (28). The Joint Statement also calls for availability of a licensed practitioner who is credentialed to maintain support of vital functions in any OB emergency, including capacity to start a cesarean delivery within 30 minutes of the decision to perform it (28). Inadequate physician supervision is an important cause of adverse events around delivery. One review of maternal deaths and near misses found mismanagement of patient or failure or delay in diagnosis as factors in 90% of cases (29). Studies focused on the potential for hemodynamic instability around delivery point out the structures required to manage them (28) (30). Pasupathy and colleagues found an increase from 4.2 to 5.6 neonatal deaths per 10,000 deliveries for deliveries occurring at off hours in Scotland 1985-2004. They were able to attribute much of this difference to a 1.7 fold increase in the odds of anoxia in this study of more than 1 million live births. Their conclusion was that staffing patterns led to delays which led to intrapartum anoxia which led to death. (Pasupathy D, Wood AM, Pell JP, et al. Time of Birth and Risk of Neonatal Death at Term: Retrospective Cohort Study. BMJ. 2010. 341:c334-349). Despite the absence of a randomized trial of 24/7 in house physician coverage of OB, the accumulated evidence supports our Expert Panel's judgment: this is a critical structural element for HROB.

Although 24-hour in-house anesthesia coverage has not been evaluated in a randomized trial, evidence suggests that inadequate anesthesiologist supervision is associated with maternal death. In a study of 18 years of anesthesia-related deaths in Michigan, more than half were attributed to inadequate supervision by an anesthesiologist (7). Adequate monitoring by an anesthesiologist is vital, as nearly one-third of all births in the U.S. are cesarean deliveries, an increase of nearly 50% since 1996 (31). Risk of death for women with an emergency cesarean section is 3 times as high as those with a planned cesarean section, suggesting that physicians capable of safely performing an emergent cesarean section is of great concern in obstetric care. Moreover, in a retrospective study examining 1.5 million deliveries from 2000 to 2006, the rate of maternal mortality was 10-fold higher with Cesarean delivery compared with a vaginal mode of delivery (32). Consistent with well-documented increases in maternal mortality in cesarean versus vaginal deliveries, the risk of severe maternal morbidity also increases (5-10 times higher), which includes hemorrhage and increased blood loss (33). Pregnancy-related hemorrhage and transfusion rates have increased substantially over the last decade (4). Therefore, on-site blood banking/transfusion services are imperative for planned and emergency cesareans, as well as other complications resulting in hemorrhage and extensive maternal blood loss. In their program that designates Regional Perinatal Centers, our partners in the New York State Department of Health operationalize it as: "24-hour capability to provide blood group, Rh Type, cross-matching, antibody testing…Either ABO specific or 0-Rh-negative blood and fresh frozen plasma and cryoprecipitate available at the facility at all times…"

Postpartum hemorrhage remains one of the most significant maternal complications of childbirth in the United States, with peripartum transfusion the most commonly identified morbidity (34). Given the increased risk for transfusion among women with anemia and placentation disorders, we will assess on-site blood banking for all high-risk deliveries (35).

Lastly, our definition of high-risk deliveries includes deliveries of low birthweight infants. There is an abundance of literature that has demonstrated that very small infants delivered in level 3 nurseries have better outcomes (23, 36). Further evidence is cited above, as is the AAP /ACOG Guideline on Perinatal Care. In the 1970's, regionalization of perinatal care was instituted in the United States and evaluations have demonstrated that antepartum risk identification and transfer of management of high-risk pregnancies to tertiary centers for delivery resulted in reduced neonatal mortality (24). Regionalization of perinatal care has been widely accepted in the United States and reaffirmed in a recent

American Academy of Pediatrics Policy Statement (12). We include a broader definition of high risk and now propose to measure the proportion of high-risk deliveries that occur in hospitals with Level III or higher neonatal intensive care units.

1a.8.2. Provide the citation and summary for each piece of evidence.

Please see Appendices for both Expert Panel Criteria and Literature Summary.

1. Evidence, Performance Gap, Priority – Importance to Measure and Report

Extent to which the specific measure focus is evidence-based, important to making significant gains in healthcare quality, and improving health outcomes for a specific high-priority (high-impact) aspect of healthcare where there is variation in or overall less-than-optimal performance. ***Measures must be judged to meet all subcriteria to pass this criterion and be evaluated against the remaining criteria.***

1a. Evidence to Support the Measure Focus – See attached Evidence Submission Form
Facility_evidence_attachment_submission_3_22_16-b.docx

1b. Performance Gap

Demonstration of quality problems and opportunity for improvement, i.e., data demonstrating:

- considerable variation, or overall less-than-optimal performance, in the quality of care across providers; and/or
- disparities in care across population groups.

1b.1. **Briefly explain the rationale for this measure** *(e.g., the benefits or improvements in quality envisioned by use of this measure)*
The health of children in the United States is fostered by healthy pregnancies and deliveries that produce healthy mothers and healthy babies. The availability of high-risk obstetric services are critical for the health of pregnant women with high-risk deliveries and ultimately for the health of their unborn infant. It is important to capture the extent to which women with risk factors for a complicated delivery are delivered at hospitals that provide sufficient care for safe monitoring of emergent cesarean sections, obstetrical anesthesia services, in-house blood banking/transfusion services, as well as a level 3 NICU.

This innovative measure addresses a complex and critical idea: How available are important high-risk obstetrical (HROB) services to women who may need them? Based upon a national expert panel's recommendations and consistent with emerging constructs from ACOG and other leading women's health organizations, we set forth specifications to identify pregnancies that constitute high risk. We assess four critical structural elements to assess their presence or absence (onsite and round the clock) in the facilities in which high risk women deliver. The practices are: coverage of the OB service by a physician capable of managing labor and delivery and performing an emergent C-section; dedicated coverage of the OB service by an anesthesiologist qualified to provide OB anesthesia, transfusion services; and a Level 3 or higher NICU.

This submission specifically responds to an assignment from AHRQ and CMS to develop measures of the availability of high risk obstetrical care as a Center of Excellence in the Pediatric Quality Measurement Program. We have used a rigorous and systematic process that was highly engaged with clinicians, stakeholders, and experts to develop these measures. We began with the evidence base and the literature.

High-risk women suffer increased rates of maternal or infant morbidity and mortality. Maternal deaths and near misses are often preventable through improved quality and safety of maternity care. The rapidly rising rate of cesarean sections and associated complications points out the need for OB staffing by physicians. High maternal hemorrhage rates point out the critical importance of transfusion and blood bank services. And the value of NICU care for infants is well established. These are important measures regarding quality and patient safety. Racial/ethnic disparities in practice are well-documented -- these four availability measures address important gaps in quality and safety and have the potential to narrow disparities in maternal and neonatal outcomes. Our analysis in New York City found that one-third of the black-white disparity in infant mortality could be accounted for by the hospital of birth (Howell et al, Pediatrics).

These were designed to be constructs that function as a population index; as population measures, we have tested them. As intended, our validation tests showed that more geographically isolated areas show less availability than areas with more dense medical services. We found the four constructs are complementary and not duplicative. They were sensitive to differences in socioeconomic status, race, and urbanicity. We found they could be implemented in New York State Medicaid data. Class A determination of eligibility was readily accomplished for those states that fully populated the MAX Medicaid data base. The measures performed well in both sets.

In 2010 in NY State Medicaid, 24.52% of Class A deliveries, 27.98% of Class B, and 24.66% of unduplicated combined A and B occurred in hospitals that met the structural measure for OB care.

1b.2. **Provide performance scores on the measure as specified (current and over time) at the specified level of analysis.** *(This is required for endorsement maintenance. Include mean, std dev, min, max, interquartile range, scores by decile. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included).*

This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.
Using AHA data to categorize hospitals. Shown both with missing data as the absence of Level 3 OB care (Blue) and excluding those with missing data. Other than LA and IL, most states show similar rates. Data is from Medicaid MAX data. (Appendix)

A similar chart with data regarding Level 3 nurseries is shown in the Appendix as well.

These charts indicate that large numbers of high risk pregnancies are delivered across the United States in hospitals without Level 3 OB services (which by AHA definitions are not as strict as the specifications on our measure) and in hospitals without at least Level 3 NICU.

1b.3. If no or limited performance data on the measure as specified is reported in 1b2, then provide a summary of data from the literature that indicates opportunity for improvement or overall less than optimal performance on the specific focus of measurement.

1b.4. Provide disparities data from the measure as specified (current and over time) by population group, e.g., by race/ethnicity, gender, age, insurance status, socioeconomic status, and/or disability. *(This is required for endorsement maintenance. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities include.) This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*
Race/Ethnicity

Our feasibility assessment confirmed that racial and ethnicity data are almost universally available and that method of assignment of race and ethnicity to the mother varied. It could be based on maternal self-report or assigned by the hospital. For the purposes of this measure we use the existing data as recorded in the mothers' medical records.

Testing sites that participated in the CAPQuaM feasibility assessment were asked to determine if maternal race/ethnicity was documented in the maternal chart, the infant chart, or if the information was located in both charts. Sites were also asked if infant race/ethnicity was documented in the maternal chart, the infant chart, or both charts. Representatives from institutions were asked to determine whether the data source for maternal race/ethnicity was located in an electronic medical record format (EMR) or a paper format. Institutions were also asked to indicate the difficulty of data abstraction in obtaining maternal race/ethnicity. Responses included very difficult to collect, difficult to collect, not difficult to collect, or unavailable. Virtually all indicated that this was not difficult to collect. The data was generally on the electronic medical records. The New York State Medicaid Program was able to identify race using their information systems. Forty-five individuals out of nearly 60,000 pregnancies were missing data on race.

We also examined race/ethnicity data in New York State Medicaid files. The following statistics focus on women found to be high risk by at least one of our two approaches identifying high-risk women. Although the scarcity of black women having babies in rural counties limited the scope of our analyses, we were able to see racial differences in the more urban counties. In Large Metropolitan areas among women who met our criteria for high-risk deliveries, 44.76% of black women, 40.11% of Hispanic women, and 30.04% of white women in Medicaid delivered in hospitals with Level 3 or higher NICUs. This may reflect housing patterns with increased numbers of minorities in inner cities, more proximate to hospitals with these services. This hypothesis is supported because those living in smaller metropolitan areas (under 250,000), show both lower rates and a different distribution: black women at 33.54%, white women at 20.04%, and Hispanic women at 13.89%.

A different pattern is seen with regional perinatal centers, our proxy for 24/7 blood banking/transfusion centers. For large metro areas among women who met our criteria for high-risk deliveries, 19.25% of white women, 13.92% of black women, and 13.82% of Hispanic women deliver at these institutions. Still a slightly different pattern (black>Hispanic>white) is seen in large metro areas for our 24/7in house OB proxy measure.

We found that our measures are able to identify statistically significant differences in performance across race/ethnicity, poverty, and also when stratifying for several of the levels of urbanicity.

Socioeconomic Status

Institutions participating in feasibility assessments were asked to determine whether sources of payment could be found in patient charts. Payment sources were identified as being in the form of an electronic medical record (EMR) or a paper record. Representatives from the participating institutions were then asked to assess the difficulty of data abstraction of the payment source. The data was generally on the electronic medical records. Our feasibility testing demonstrated that we can use Medicaid

insurance as a marker for SES and our New York City data demonstrate this to be an important independent predictor of poor maternal and infant outcomes.

We further use the national distribution of percent of individuals in poverty to establish five categories that reflect the counties level of poverty. We considered other data such as county median income or county unemployment, but felt that the percent of individuals in poverty was a more integrative measure. The use of a geographic rather than an individual measure is consistent with recent applications of hierarchical methods to study the impact of poverty and also with data that indicate that local disparities in income are an independent predictor of outcomes. It also allows this measure to consider issues of socioeconomic status while using publicly available data and requiring only the mother's county of residence, a more reliable data point than self- reported income.

Our analysis of USDA data considering 3142 counties and related geographic units found a mean of 17.2 % of county residents living in poverty, a standard deviation of 6.5%, and an interquartile range of 8.2%. The distribution illustrated below, shows meaningful dispersion and supports our plan to build off quartiles of distribution with a finer focus in higher areas of poverty. See Table 7 below.

Table 7

Quantile	Percent in Poverty
Maximum	49.9%
99	37.5%
95	28.9%
90	25.7%
75	20.7%
50	16.5%
25	12.5%
10	10.0%
5	8.6%
1	6.1%
Minimum	2.9%

All of New York State lies in the top three quartiles. We would expect to find the largest differences between poorer and other counties than across the upper end of the spectrum. Nonetheless, we conducted the analysis and found statistically significant differences. Quartile 2 was slightly better than the top quartile in performance, but the third quartile, below the median, had less than half the proportion of high-risk women delivering at sites with each of the structural attributes than Quartile 2. See Table 8 below. Interestingly, poor counties performed better than did the most rural counties, confirming that these various approaches to stratification are capturing different information.

Table 8
HROB Summary (Combined Unduplicated)
New York State Medicaid, 2010

Poverty Level	N	OB Proxy*	Transfusion Proxy **	NICU >=3***
Top Quartile	8,533	20.27%	12.83%	25.08%
Second Quartile	44,013	26.58%	14.42%	37.89%
MEDIAN INCOME				
Third Quartile	3,919	10.56%	2.91%	9.95%

* Proxy for OB 24/7 coverage is Level 3 Obstetrical Care according to the American Hospital Association Survey.
** Proxy for 24/7 Transfusion services is designation s a Regional Perinatal center by New York State, for which 24/7 transfusion services are a requirement.
*** NICU represents provision of Level 3 or Level 4 NICU services as identified by frequent the billing of Medicaid for Level 3 or Level 4 NICU services using Revenue Codes 173 or 174.

Rurality/Urbanicity

As described in the specification, we used the Urban Influence Codes (UIC) below to describe the level of rurality or urbanicity.

Metropolitan
1. In large metro area of 1+ million residents
2. In small metro area of less than 1 million residents
3. Micropolitan adjacent to large metro
4. Non-core adjacent to large metro
5. Micropolitan adjacent to small metro
6. Non-core adjacent to small metro with own town
7. Non-core adjacent to small metro no own town
8. Micropolitan not adjacent to a metro area
9. Non-core adjacent to micro with own town
10. Non-core adjacent to micro with no own town
11. Non-core not adjacent to metro or micro with own town
12. Non-core not adjacent to metro or micro with no own town

We analyzed 3143 county equivalents in the U.S, and the results are shown in Table 9 below.

Table 9

UIC_2013

UIC_2013	Frequency	Percent
1	432	13.74
2	735	23.39
3	130	4.14
4	149	4.74
5	242	7.70
6	344	10.94
7	162	5.15
8	269	8.56
9	184	5.85
10	189	6.01
11	125	3.98
12	182	5.79

The population is heavily weighted to metropolitan areas as demonstrated in Table 10 below.

Table 10
UIC_2013

UIC_2013	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	1.672E8	55.07	1.672E8	55.07
2	91886000	30.27	2.5909E8	85.34
3	6921700	2.28	2.6601E8	87.62
4	3094100	1.02	2.691E8	88.54
5	10760300	3.54	2.7986E8	92.18
6	7005400	2.31	2.8687E8	94.49
7	1511900	0.50	2.8838E8	94.99
8	8459500	2.79	2.9684E8	97.78
9	2684400	0.88	2.9952E8	98.66
10	1289100	0.42	3.0081E8	99.09
11	1887800	0.62	3.027E8	99.71
12	887700	0.29	3.0359E8	100.00

As noted, we use Urban Influence Codes (UIC), which have been developed by the USDA based on a number of criteria to describe the levels of urbanicity and rurality. This is intended not only to report within plan differences but to allow for aggregation as appropriate. While each UIC has its own meaningful definition, some researchers choose to aggregate various codes. We recommend consideration of the aggregation schema of Bennett and colleagues at the South Carolina Rural Research Center (44). Their aggregation scheme brings together Codes 1 & 2 as Urban; 3, 5, & 8 as micropolitan rural; 4, 6, & 7 as rural adjacent to a metro area; and 9, 10, 11, & 12 as remote rural. We observe that UIC 5 might also be aggregated with 4, 6, & 7 as an adjacent rural area. Further,

this approach to rurality does not map exactly to the population density based definition of frontier (< 6 persons per square mile) as articulated in the Affordable Care Act. However, use of such categories is consistent with the ACA's intent that the Secretary ask that data collected for racial and ethnic disparities also look at underserved frontier counties. Frontier health care may be approximated by analysis of the remote rural categories (45). Our judgment was confirmed after CAPQuaM consulted with Gary Hart, Director of the Center for Rural Health at the University of North Dakota School of Medicine & Health Sciences, who is heading a HRSA-funded project to develop new methods to analyze frontier health. We clarified that his work suggests that UIC 9-12 is the best overall approach to using county level data to study frontier health. Inclusion of UIC 8 would make the analysis more sensitive to including frontier areas but at a meaningful cost in specificity.

Those interested in care specific to large cities may wish to aggregate rural areas and analyze UIC 1 and 2 separately.

The New York State Medicaid data was sensitive to urbanicity and this is described elsewhere.

1b.5. **If no or limited data on disparities from the measure as specified is reported in 1b4, then provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement. Include citations.**

1c. High Priority (previously referred to as High Impact)

The measure addresses:

- a specific national health goal/priority identified by DHHS or the National Priorities Partnership convened by NQF; OR
- a demonstrated high-priority (high-impact) aspect of healthcare (e.g., affects large numbers of patients and/or has a substantial impact for a smaller population; leading cause of morbidity/mortality; high resource use (current and/or future); severity of illness; and severity of patient/societal consequences of poor quality).

1c.1. **Demonstrated high priority aspect of healthcare**

Affects large numbers, A leading cause of morbidity/mortality, Frequently performed procedure, High resource use, Patient/societal consequences of poor quality

1c.2. **If Other:**

1c.3. **Provide epidemiologic or resource use data that demonstrates the measure addresses a high priority aspect of healthcare. List citations in 1c.4.**

Childbirth is common and universally experienced. Our analysis of Mediciad data shows that childbirth accounts for a plurality of hospital admissions for Medicaid programs, so this measure has special importance to state health programs. Our data show that state to state between one- and two-thirds of these admissions are high risk. Hospital costs for childbirth and neonatal care are large.

1a.3. provides additional epidemiologic and resource use data that demonstrates the measure addresses a high priority aspect of healthcare.

1c.4. **Citations for data demonstrating high priority provided in 1a.3**

1. Hall SA, Kaufman JS, Ricketts TC. Defining urban and rural areas in U.S. epidemiologic studies. J Urban Health. 2006;83(2):162-75. PMCID: 2527174.
2. Hankins GD, Clark SL, Pacheco LD, O´Keeffe D, D´Alton M, Saade GR. Maternal mortality, near misses, and severe morbidity: lowering rates through designated levels of maternity care. Obstet Gynecol. 2012;120(4):929-34.
3. Berg CJ, Callaghan WM, Syverson C, Henderson Z. Pregnancy-related mortality in the United States, 1998 to 2005. Obstet Gynecol. 2010;116(6):1302-9.
4. Callaghan WM, Creanga AA, Kuklina EV. Severe maternal morbidity among delivery and postpartum hospitalizations in the United States. Obstet Gynecol. 2012;120(5):1029-36.
5. Howell EA, Stone J, Kleinman LC, Inamdar S, Matseoane S, Chassin MR. Approaching NIH guideline recommended care for maternal-infant health: clinical failures to use recommended antenatal corticosteroids. Matern Child Health J. 2010;14(3):430-6.
6. Howell EA, Holzman I, Kleinman LC, Wang J, Chassin MR. Surfactant use for premature infants with respiratory distress syndrome in three New York city hospitals: discordance of practice from a community clinician consensus standard. J

Perinatol. 2010;30(9):590-5. PMCID: 2888640.
7. Howell EA, Hebert P, Chatterjee S, Kleinman LC, Chassin MR. Black/white differences in very low birth weight neonatal mortality rates among New York City hospitals. Pediatrics. 2008;121(3):e407-15.
8. Hoyert DL, Danel I, Tully P. Maternal mortality, United States and Canada, 1982-1997. Birth. 2000;27(1):4-11.
9. Berg CJ, Atrash HK, Koonin LM, Tucker M. Pregnancy-related mortality in the United States, 1987-1990. Obstet Gynecol. 1996;88(2):161-7.
10. Nannini A, Weiss J, Goldstein R, Fogerty S. Pregnancy-associated mortality at the end of the twentieth century: Massachusetts, 1990-1999. J Am Med Womens Assoc. 2002;57(3):140-3.
11. Commission TJ. Preventing Maternal Death: Sentinel Event Alert; 2010 January 26, 2010 Contract No.: Document Number|.
12. Levels of neonatal care. Pediatrics. 2012;130(3):587-97.
13. Collaborative CMQC. OB Hemorrhage Toolkit. 2010 [updated 2010; cited]; Available from: http://cmqcc.org/ob_hemorrhage.
14. DeFrances CJ, Cullen KA, Kozak LJ. National Hospital Discharge Survey: 2005 annual summary with detailed diagnosis and procedure data. Vital Health Stat 13. 2007(165):1-209.
15. Andrews R. The national hospital bill: the most expensive conditions by payer. HCUP Statistical Brief 59. Rockville, MD: Agency for Healthcare, Research and Quality; 2008.
16. Health NYSDo. Table 4: Live Birth Summary by Mother´s Race/Ethnicity, New York State 2011. 2013 [updated 2013; cited]; Available from: http://www.health.ny.gov/statistics/vital_statistics/2011/table04.htm.
17. Association AH. Chart Book: Prepared to Care the 24/7 Role of America´s Full-service Hospitals; 2005 Contract No.: Document Number|.
18. Health NYSDo. New York State Hospital Profile 2012 [updated 2012; cited]; Available from: http://hospitals.nyhealth.gov/index.php.
19. Committee Opinion no. 514: emergent therapy for acute-onset, severe hypertension with preeclampsia or eclampsia. Obstet Gynecol. 2011;118(6):1465-8.
20. Bucklin BA, Hawkins JL, Anderson JR, Ullrich FA. Obstetric anesthesia workforce survey: twenty-year update. Anesthesiology. 2005;103(3):645-53.
21. ACOG committee opinion No. 433: optimal goals for anesthesia care in obstetrics. Obstet Gynecol. 2009;113(5):1197-9.
22. Barbieri R, Camann WR. A pivotal partnership: obstetrics and anesthesiology [Editorial]. OB Management. 2002(March):8-9.
23. Phibbs CS, Bronstein JM, Buxton E, Phibbs RH. The effects of patient volume and level of care at the hospital of birth on neonatal mortality. JAMA. 1996;276(13):1054-9.
24. McCormick MC, Shapiro S, Starfield BH. The regionalization of perinatal services. Summary of the evaluation of a national demonstration program. JAMA. 1985;253(6):799-804.
25. Aday LA, Andersen R. A framework for the study of access to medical care. Health Serv Res. 1974;9(3):208-20. PMCID: 1071804.
26. Andersen RM, McCutcheon A, Aday LA, Chiu GY, Bell R. Exploring dimensions of access to medical care. Health Serv Res. 1983;18(1):49-74. PMCID: 1068709.
27. Kuhlthau KA. Measures of availability of health care services for children. Acad Pediatr. 2011;11(3 Suppl):S42-8.
28. ACOG Committee Opinion No. 433: Optimal Goals for Anesthesia Care in Obstetrics; 2009 Contract No.: Document Number|.
29. Geller SE, Rosenberg D, Cox SM, Brown ML, Simonson L, Driscoll CA, et al. The continuum of maternal morbidity and mortality: factors associated with severity. Am J Obstet Gynecol. 2004;191(3):939-44.
30. Pasupathy D, Wood AM, Pell JP, Fleming M, Smith GC. Time of birth and risk of neonatal death at term: retrospective cohort study. BMJ. 2010;341:c3498.

PMCID: 2904877.
31. Martin JA, Hamilton, B.E., Ventura S.J., Osterman M, WIlson E.C., Mathews T.J. Births: Final Data for 2010. National Center for Health Statistics; August 28, 2012 Contract No.: Document Number|.
32. Clark SL, Belfort MA, Dildy GA, Herbst MA, Meyers JA, Hankins GD. Maternal death in the 21st century: causes, prevention, and relationship to cesarean delivery. Am J Obstet Gynecol. 2008;199(1):36 e1-5; discussion 91-2 e7-11.
33. Shearer EL. Cesarean section: medical benefits and costs. Soc Sci Med. 1993;37(10):1223-31.
34. Ehrenthal DB, Chichester ML, Cole OS, Jiang X. Maternal risk factors for peripartum transfusion. J Womens Health (Larchmt). 2012;21(7):792-7.
35. Rouse DJ, MacPherson C, Landon M, Varner MW, Leveno KJ, Moawad AH, et al. Blood transfusion and cesarean delivery. Obstet Gynecol. 2006;108(4):891-7.
36. Chung JH, Phibbs CS, Boscardin WJ, Kominski GF, Ortega AN, Needleman J. The effect of neonatal intensive care level and hospital volume on mortality of very low birth weight infants. Med Care. 2010;48(7):635-44.
37. Medicine Io. Crossing the Quality Chasm: The IOM Health Care Quality Initiative, Contract No.: Document Number|.
38. Kuklina EV, Whiteman MK, Hillis SD, Jamieson DJ, Meikle SF, Posner SF, et al. An enhanced method for identifying obstetric deliveries: implications for estimating maternal morbidity. Matern Child Health J. 2008;12(4):469-77.
39. Diamond CRK, Kohler SA. . Use of Paper Medical Records Versus Administrative Data for Measuring and Improving Health Care Quality: Are We Still Searching for a Gold Standard? Disease Management. 2001;4(3):121-30.
40. Hospitals AASo. Data Collection Methods. Health Forum, LLC; [cited]; Available from: http://www.ahadataviewer.com/about/data/.
41. Virnig BA, McBean M. Administrative data for public health surveillance and planning. Annu Rev Public Health. 2001;22:213-30.
42. Mangione-Smith R, DeCristofaro AH, Setodji CM, Keesey J, Klein DJ, Adams JL, et al. The quality of ambulatory care delivered to children in the United States. N Engl J Med. 2007;357(15):1515-23.
43. Rubio DM B-WM, Tebb SS, Lee ES, Rauch S. Objectifying content validity: Conducting a content validity study in social work research. Social Work Research. 2003;27(2):94-104.
44. Bennett KJ, Olatosi B. & Probst, J.C. Health Disparities: A rural-urban chartbook. Columbia, South Carolina: South Carolina Rural Health Research Center; 2008.
45. Hart G. Frontier/Remote Island, and Rural Literature Review2012: Available from: http://ruralhealth.und.edu/frontier/pdf/lit_review.pdf.

1c.5. If a PRO-PM (e.g. HRQoL/functional status, symptom/burden, experience with care, health-related behaviors), provide evidence that the target population values the measured PRO and finds it meaningful. *(Describe how and from whom their input was obtained.)*

n/a

1d. Composite Quality Construct and Rationale

1d.1. A composite performance measure is a combination of two or more component measures, each of which individually reflects quality of care, into a single performance measure with a single score.

For purposes of NQF measure submission, evaluation, and endorsement, the following will be considered composites:

- Measures with two or more individual performance measure scores combined into one score for an accountable entity.
- Measures with two or more individual component measures assessed separately for each patient and then aggregated into one score for an accountable entity:
    - all-or-none measures (e.g., all essential care processes received, or outcomes experienced, by each patient); or
    - any-or-none measures (e.g., any or none of a list of adverse outcomes experienced, or inappropriate or unnecessary care processes received, by each patient).

1d.1. Please identify the composite measure construction: all-or-none measures (e.g., all essential care processes received, or outcomes experienced, by each patient)

1d.2. Describe the quality construct, including:

- the overall area of quality
- included component measures and
- the relationship of the component measures to the overall composite and to each other.

The quality construct is the availability of High Risk Obstetrical Care.
This measure is comprised of four sub constructs, each of which refers to a strongly recommended structural aspect of the facilities in which women deliver infants:
1. 24/7 presence of a physician in house who is skilled at managing labor and is capable of performing an emergent C-section.
2. 24/7 presence of an anesthesiologist skilled in providing obstetrical anesthesia;
3. 24/7 presence of blood banking services capable of type and cross and providing rapid access to transfusion
4. Level 3 or higher neonatal intensive care unit (NICU)

Normatively, these composites should move together with Centers that deliver high risk infants having all of these. In our assessment we did not find that they moved in unison and there is additive information. Patient safety is best supported by having each of the services on site at all times for deliveries that are above normal risk

1d.3. Describe the rationale for constructing a composite measure, including how the composite provides a distinctive or additive value over the component measures individually.
Each of the components has independent information that may inform improvement. As an accountability measures, this is an all or none composite.

1d.4. Describe how the aggregation and weighting of the component measures are consistent with the stated quality construct and rationale.
As an all or none measure, there are no weighting schemas required. Each of the components represents an important capacity for facilities that deliver high risk pregnancies, and patient safety is served best when each is present.

2. Reliability and Validity—Scientific Acceptability of Measure Properties

Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. ***Measures must be judged to meet the subcriteria for both reliability and validity to pass this criterion and be evaluated against the remaining criteria.***

2a.1. Specifications The measure is well defined and precisely specified so it can be implemented consistently within and across organizations and allows for comparability. eMeasures should be specified in the Health Quality Measures Format (HQMF) and the Quality Data Model (QDM).

De.5. **Subject/Topic Area** *(check all the areas that apply):*
Perinatal and Reproductive Health, Perinatal and Reproductive Health : Newborn, Perinatal and Reproductive Health : Perinatal, Prevention, Surgery : Perioperative

De.6. **Cross Cutting Areas** *(check all the areas that apply):*
Access, Disparities, Prevention, Safety, Safety : Complications

S.1. **Measure-specific Web Page** *(Provide a URL link to a web page specific for this measure that contains current detailed specifications including code lists, risk model details, and supplemental materials. Do not enter a URL linking to a home page or to general information.)*
n/a

S.2a. **If this is an eMeasure**, HQMF specifications must be attached. Attach the zipped output from the eMeasure authoring tool (MAT) - if the MAT was not used, contact staff. (Use the specification fields in this online form for the plain-language description of the specifications)
This is not an eMeasure **Attachment:**

S.2b. **Data Dictionary, Code Table, or Value Sets** *(and risk model codes and coefficients when applicable) must be attached. (Excel or csv file in the suggested format preferred - if not, contact staff)*
Attachment **Attachment:** Structural_Attributes_Data_Dictionary_submission.xlsx

S.3. **For endorsement maintenance**, please briefly describe any changes to the measure specifications since last endorsement date and explain the reasons.
n/a

S.4. **Numerator Statement** *(Brief, narrative description of the measure focus or what is being measured about the target population, i.e., cases from the target population with the target process, condition, event, or outcome)*
IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
Number of eligible newborn deliveries that occur in facilities with:
(a) Level 3 or higher NICU services on campus. Level 3 NICU is defined as meeting either the American Academy of Pediatrics (AAP) criteria or a locally used set of explicit criteria recognized by that state's Department of Health.
(b) 24/7 on-site blood banking services/transfusion services that are always available for obstetrical patients. By 24/7 blood banking/transfusion services we mean that the following are always available to obstetrical patients: testing of blood group and Rh Type; cross matching; antibody testing; transfusion with on-site and available blood, either ABO specified or O-Rh-negative; transfusion with fresh frozen plasma; and transfusion with cryoprecipitate.
(c) 24/7 in - house physician dedicated to labor and delivery who is capable of safely managing labor and delivery, and of performing a cesarean section, including an emergent cesarean section.
(d) 24/7 in - house physician coverage dedicated to the obstetrical service by an anesthesiologist who is qualified to provide obstetrical anesthesia.

Measure: Meets all four criteria.

Stratifications:
a. Meets none
b. Includes a
c. Includes b
d. includes c

e. includes d

Numerator Elements:
Number of eligible deliveries
Maternal and infant ICD-9 codes
Response to survey question identified on technical specifications or Other valid self-report of structural characteristics as specified

No Numerator Exclusions

S.5. **Time Period for Data** *(What is the time period in which data will be aggregated for the measure, e.g., 12 mo, 3 years, look back to August for flu vaccination? Note if there are different time periods for the numerator and denominator.)*
The Reporting Period is one year. Preferred look back period is 2 years before delivery.

S.6. **Numerator Details** *(All information required to identify and calculate the cases from the target population with the target process, condition, event, or outcome such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
IF an OUTCOME MEASURE, describe how the observed outcome is identified/counted. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
Of those denominator births, Report as numerator the number that were born in facilities that report meeting all four of the listed criteria.

(a) Level 3 or higher NICU services on campus. Level 3 NICU is defined as meeting either the American Academy of Pediatrics (AAP) criteria or a locally used set of explicit criteria recognized by that state's Department of Health.
(b) 24/7 on-site blood banking services/transfusion services that are always available for obstetrical patients. By 24/7 blood banking/transfusion services we mean that the following are always available to obstetrical patients: testing of blood group and Rh Type; cross matching; antibody testing; transfusion with on-site and available blood, either ABO specified or O-Rh-negative; transfusion with fresh frozen plasma; and transfusion with cryoprecipitate.
(c) 24/7 in - house physician dedicated to labor and delivery who is capable of safely managing labor and delivery, and of performing a cesarean section, including an emergent cesarean section.
(d) 24/7 in - house physician coverage dedicated to the obstetrical service by an anesthesiologist who is qualified to provide obstetrical anesthesia.

Maintain a separate count of those who meet each of the four criteria for stratification purposes.

S.7. **Denominator Statement** *(Brief, narrative description of the target population being measured)*
Overall number of newborn deliveries in health care facilities that are born to women whose pregnancy meets the criteria for high risk. While qualification for the denominator requires that the birth occur in a health care facility this measure is not specified to assess performance of individual facilities.

S.8. **Target Population Category** *(Check all the populations for which the measure is specified and tested if any):*
Children's Health, Maternal Health, Populations at Risk, Populations at Risk : Individuals with multiple chronic conditions

S.9. **Denominator Details** *(All information required to identify and calculate the target population/denominator such as definitions, specific data collection items/responses , code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
Eligible deliveries are identified in two distinct ways. Maternal and infant ICD-10 codes are specified in S.2b. Data Dictionary Code Table.

1. Class A: Maternal Diagnoses and Comorbidities
2. Class B: Delivery Complications, Fetal Injury or Compromise, or Suboptimal Infant Diagnoses
a.Maternal Delivery Complication Codes (ICD9)
b.Maternal Stillbirth or Birth Hypoxia/Asphyxia Codes
c.Premature or small infant (Infant codes)
3. Either Class A or Class B (Unduplicated union of Class A and Class B)

The Unduplicated Union is the population eligible for inclusion in this measure. If the reporting entity is unable to link maternal and

infant records, then Class A and Class B should be reported distinctly. These data can be aggregated within plans and states; or compare performance of plans and states

Denominator Elements:
Number of deliveries
Maternal and infant ICD-9 codes
Maternal DRG, CPT codes, and revenue codes when available

DATA DICTIONARY TABLE OF CONTENTS (S2b)

- Tab1. Inclusion – Identification of Deliveries of Interest. This includes revenue codes, outcome of delivery, diagnosis-related group (DRG) delivery codes, and selected delivery related procedures.
- Tab2. Exclusions
- Tab3. Identification of women in need of high-risk services – Class A: Maternal Diagnoses and Comorbidities
- Tab4. Identification of women in need of high-risk services – Class B: Delivery Complications, Fetal Risk or Compromise, or Suboptimal Infant Outcomes

S.10. **Denominator Exclusions** *(Brief narrative description of exclusions from the target population)*
None

S.11. **Denominator Exclusion Details** *(All information required to identify and calculate exclusions from the denominator such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
n/a

S.12. **Stratification Details/Variables** *(All information required to stratify the measure results including the stratification variables, definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format with at S.2b)*
Stratifications:

a. Meets none
b. Includes a
c. Includes b
d. includes c
e. includes d

Within each stratification report the Measure for the Unduplicated Union of Class A and Class B as well as those who meet Class A and those who meet Class B.

S.13. **Risk Adjustment Type** (Select type. Provide specifications for risk stratification in S.12 and for statistical model in S.14-15)
No risk adjustment or risk stratification
If other:

S.14. **Identify the statistical risk model method and variables** *(Name the statistical method - e.g., logistic regression and list all the risk factor variables. Note - risk model development and testing should be addressed with measure testing under Scientific Acceptability)*
n/a

S.15. **Detailed risk model specifications** *(must be in attached data dictionary/code list Excel or csv file. Also indicate if available at measure-specific URL identified in S.1.)*
Note: Risk model details (including coefficients, equations, codes with descriptors, definitions), should be provided on a separate worksheet in the suggested format in the Excel or csv file with data dictionary/code lists at S.2b.

S.15a. **Detailed risk model specifications** (*if not provided in excel or csv file at S.2b*)
n/a

S.16. **Type of score:**
Rate/proportion

If other:

S.17. **Interpretation of Score** *(Classifies interpretation of score according to whether better quality is associated with a higher score, a lower score, a score falling within a defined interval, or a passing score)*
Better quality = Higher score

S.18. **Calculation Algorithm/Measure Logic** *(Describe the calculation of the measure score as an ordered sequence of steps including identifying the target population; exclusions; cases meeting the target process, condition, event, or outcome; aggregating data; risk adjustment; etc.)*
Step 1: Identify all eligible deliveries that occurred in medical facilities, using the criteria above. Eligibility can be defined on the basis of affiliation with a health plan at he time of birth, or by geographic entity as appropriate for the specified accountability entity.

Step 2: Link Maternal and Infant charts

Step 3: Identify Class A, Class B, and an unduplicated list that represents the union of Class A and Class B high-risk pregnancies. These are Denominator 1, Denominator 2, and Denominator 3 respectively. These are considered women in potential need of high-risk services ("high-risk" for short).

Step 4: Identify each health care facility that has at least one delivery that is in Denominator 1 or Denominator 2

Step 5: Identify which of those health care facilities answered "Yes" to each of the four self-report questions (or have reported elsewhere that they meet these criteria in full). Classify each facility by whether or not they answered "Yes" to all four and to each one (The latter for stratification purposes). Missing data regarding the facility are considered to be "No" responses. For the purposes of this measure calculation consider as Class 1 facilities answering "Yes" to all four questions; consider as Class 2 facilities those that did not answer "Yes" four. For stratification purposes, Class 1 facilities are those that responded yes to the specific criterion for that stratum. Those that answer "No" to all four questions are the numerator for stratum (a).

Step 6: Collect the following data elements for all eligible women
•Race
•Ethnicity
•Zip Code, State and County or equivalent area of mother's residence. Record FIPS if available

Step 7: At option of accountability entity, create stratification variables as follows:
i.Race/Ethnicity: Hispanic, Non-Hispanic Black, Non-Hispanic White; Non-Hispanic Asian/Pacific Islander, Other Non-Hispanic
ii.Urban Influence Code. Identify the Urban Influence Code(1) or UIC. (2013 urban influence codes available at: http://www.ers.usda.gov/data-products/urban-influence- codes.aspx#.UZUvG2cVoj8). Use mother's place of residence to determine UIC. State and County names can be linked or looked up directly or zip codes can be linked to County indirectly, using the Missouri Census Data Center (http://mcdc.missouri.edu/). These data will link to County or County equivalents as used in various states
•Identify the Level of Poverty in the mother's county of residence. The percent of all residents in poverty by county or county equivalent are available from the US Department of Agriculture at http://www.ers.usda.gov/data-products/county-level-data-sets/download- data.aspx. Our stratification standards are based on 2011 US population data that we have analyzed with SAS 9.3. Using Mother's state and county of residence (or equivalent) or FIPS code, use the variable PCTPOVALL_2011 to categorize into one of 5 Strata (If needed, the Missouri Data center linked above may be used to link zip codes to county equivalents):Lowest Quartile of Poverty if percent in poverty is <=12.5%
•Second Quartile of Poverty if percent in poverty is >12.5% and <=16.5%
•Third Quartile of poverty if percent in poverty is >16.5% and <=20.7%
•First Upper Quartile (75th-90th) if percent in poverty is >20.7% and <=25.7%Second Upper Quartile (>90th percentile)
iv.Insurance type (Public/Medicaid, Private/Commercial, None, other)
v.Benefit type: HMO vs PPO vs FFS vs PCCM vs Other

Step 8: Overall, and for each stratification category, count the number of high-risk deliveries that occurred in Class 1 facilities for each of the three ways of qualifying (Class A, Class B, and Union of Class A or Class B. These are Numerator 1, Numerator 2, and Numerator 3, respectively.

Step 9: Overall, and for each stratification category, calculate the percentage of high-risk pregnancies that were delivered in Class 1 and Class 2 Hospitals.
•Percentage 1 is calculated as the 100*Numerator1/Denominator 1
•Percentage 2 is calculated as the 100*Numerator2/Denominator 2
•Percentage 3 is calculated as the 100*Numerator3/Denominator 3
•Report all percentages to 2 decimal places
•If maternal and infant records cannot be linked and an unduplicated union cannot e determined, report Percentage 1 and Percentage 2, along with the denominator N in both Class A and Class B, as well as the overall number of Births eligible in Step 1, and the percentage of those births identified via Class A and by Class B.

Step 10: Report the results of Step 9.

Step 11: Optionally at the accountability entity's request, repeat calculations specified above for each stratification category listed below, using the following data elements. Report all strata with N of at least 250.
•Race and ethnicity
•Urban Influence Code or UIC.
•Level of Poverty in the County of Residence.
•Insurance type (Public/Medicaid, Private/Commercial, None, other)
•Benefit type: HMO vs PPO vs FFS vs PCCM vs Other

Step 12: Optionally calculate 95% confidence intervals.
a.Calculate standard error as the square root of the proportion of newborns delivered in facilities with the given structure multiplied by 1—the same proportion divided by the number of deliveries.
b.Multiply the standard error by 1.96.
c.Subtract that value from the measured proportion. Report the greater of and that number as the lower bound of the 95% confidence interval.
d.Add the product from b to the measured proportion. Use the lesser of that sum or 1 as the upper bound of the 95% confidence interval.

S.19. **Calculation Algorithm/Measure Logic Diagram URL or Attachment** *(You also may provide a diagram of the Calculation Algorithm/Measure Logic described above at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
Available in attached appendix at A.1

S.20. **Sampling** *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*
IF a PRO-PM, identify whether (and how) proxy responses are allowed.
n/a

S.21. **Survey/Patient-reported data** *(If measure is based on a survey, provide instructions for conducting the survey and guidance on minimum response rate.)*
IF a PRO-PM, specify calculation of response rates to be reported with performance measure results.
This survey should be conducted by the accountability entity to all facilities that provide birthing or delivery services. One response per facility-location.

S.22. **Missing data** (specify how missing data are handled, e.g., imputation, delete case.)
Required for Composites and PRO-PMs.
Missing data are considered "no" responses.

S.23. **Data Source** *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*
If other, please describe in S.24.
Administrative claims, Healthcare Provider Survey

S.24. **Data Source or Collection Instrument** *(Identify the specific data source/data collection instrument e.g. name of database, clinical registry, collection instrument, etc.)*
IF a PRO-PM, identify the specific PROM(s); and standard methods, modes, and languages of administration.
1.Survey of hospitals, birthing centers, and other health care facilities at which eligible women have delivered
The accountable entity should conduct a survey with responses to be provided by a designee of each facility's Chief Medical Officer

or equivalent.

Classify by answer yes or no to the following questions:

Does this facility always have 24/7 in house dedicated coverage of the obstetrical service by a physician capable of safely managing labor and delivery, and performing a cesarean section, including an emergent cesarean section?
a) Yes
b) No
c) Unsure
d) Refuse
(This is Question 1 of the four question HROB questionnaire included in the Appendix.)

Does this facility always have 24/7 in-house dedicated coverage of the obstetrical service by an anesthesiologist who is qualified to provide obstetrical anesthesia?
a) Yes
b) No
c)Unsure
d)Refuse
(This is Question 2 of the four question HROB questionnaire included in the Appendix.)

Does this facility provide 24/7 on-site blood banking services/transfusions services that are always available for obstetrical patients? By 24/7 blood banking/transfusion services we mean that the following are always available to obstetrical patients: testing of blood group and Rh Type; cross-matching; antibody testing; transfusion with on site and available blood, either ABO specific or 0-Rh-negative; transfusion with fresh frozen plasma; and transfusion with cryoprecipitate.
a) Yes
b) No
c) Unsure
d) Refuse
(This is Question 3 of the four question HROB questionnaire included in the Appendix.)

Does this facility offer level 3 or higher NICU services according to either American Academy of Pediatrics (AAP) criteria or other explicit criteria recognized by your state Department of Health??
a) Yes
b) No
c) Unsure
d) Refuse/Blank
(This is Question 4 of the four question HROB questionnaire included in the Appendix.)

2.Data with billing and diagnosis codes
a.Identify Eligible population
i.Women who have deliveries in health care facilities
ii.Identify those deliveries associated with high-risk conditions
1. Maternal record: High-risk Diagnoses
2. Maternal record: Complicated Delivery
3. Maternal record: Stillbirth or Birth Asphyxia
4. Infant record: Premature or Small Infant

3. Woman's medical record
Only if needed for maternal race, ethnicity, or data regarding place of residence.

S.25. **Data Source or Collection Instrument** *(available at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
Available in attached appendix at A.1

S.26. **Level of Analysis** *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*
Health Plan, Integrated Delivery System, Population : Community, Population : County or City, Population : National, Population : Regional, Population : State

S.27. **Care Setting** *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*
Hospital/Acute Care Facility, Other
If other: Birthing Centers

S.28. **COMPOSITE Performance Measure** - Additional Specifications *(Use this section as needed for aggregation and weighting rules, or calculation of individual performance measures if not individually endorsed.)*
n/a

2a. Reliability – See attached Measure Testing Submission Form
2b. Validity – See attached Measure Testing Submission Form
Template_MeasSubm_CompositeMeasTesting_2015-09-01_3_22_16_LK.docx

Measure Number (*if previously endorsed*)**:** Click here to enter NQF number

Composite Measure Title: Structural Attributes of Facility in which High Risk Women Deliver Newborns: A PQMP Measure

Date of Submission: 3/22/2016

Composite Construction:

☐Two or more individual performance measure scores combined into one score

☒ All-or-none measures (e.g., all essential care processes received or outcomes experienced by each patient)

☐ Any-or-none measures (e.g., any or none of a list of adverse outcomes experienced, or inappropriate or unnecessary care processes received, by each patient)

Instructions: Please contact NQF staff before you begin.

- If a component measure is submitted as an individual performance measure, the non-composite measure testing form must also be completed and attached to the individual measure submission.
- Measures must be tested for all the data sources and levels of analyses that are specified. ***If there is more than one set of data specifications or more than one level of analysis, contact NQF staff*** about how to present all the testing information in one form.
- **For all composite measures, sections 1, 2a2, 2b2, 2b3, 2b5, and 2d must be completed.**
- **For composites with outcome and resource use measures**, section **2b4** also must be completed.
- If specified for **multiple data sources/sets of specificaitions** (e.g., claims and EHRs), section **2b6** also must be completed.
- Respond to all questions as instructed with answers immediately following the question. All information on testing to demonstrate meeting the subcriteria for reliability (2a2), validity (2b2-2b6), and composites (2d) must be in this form. An appendix for *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 25 pages (*incuding questions/instructions;* minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.

Note: The information provided in this form is intended to aid the Standing Committee and other stakeholders in understanding to what degree the testing results for this measure meet NQF's evaluation criteria for testing.

2a2. Reliability testing [10] demonstrates the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise. For **PRO-PMs and composite performance measures**, reliability should be demonstrated for the computed performance score.

Presence of Level 3 Obstetrical care by state. High precision is demonstrated.

State	N Obs	N	Mean	Std Dev	Lower 95% CL for Mean	Upper 95% CL for Mean
AZ	51527	51141	0.33816	0.47309	0.33406	0.34226
ID	9553	9003	0.19427	0.39566	0.18609	0.20244
IL	60073	17164	0.24715	0.43136	0.24069	0.2536
IN	36656	36433	0.25131	0.43377	0.24686	0.25576

KS	**14410**	13678	0.52881	0.49919	0.52044	0.53717
KY	**28726**	28708	0.11613	0.32039	0.11243	0.11984
LA	**37731**	1376	0.43314	0.49569	0.40693	0.45935
MO	**33287**	27404	0.15863	0.36533	0.1543	0.16295
MT	**4016**	3997	0.35352	0.47812	0.33869	0.36834
NC	**65324**	65321	0.47032	0.49912	0.4665	0.47415
NH	**3792**	3650	0.12411	0.32975	0.11341	0.13481
NJ	**14363**	13111	0.32797	0.46949	0.31993	0.33601
NM	**14849**	14038	0.31842	0.46588	0.31071	0.32613
NY	**116151**	96499	0.16376	0.37006	0.16143	0.1661
VA	**31355**	30701	0.35015	0.47703	0.34482	0.35549
VT	**2612**	2361	0.37315	0.48374	0.35362	0.39267
WI	**23796**	18965	0.14854	0.35564	0.14347	0.1536
WY	**3242**	3239	0.046	0.20952	0.03878	0.05322

Similarly the standard error for the NICU presence is typically less than 1% with similarly narrow confidence intervals.

2b2. Validity testing [11] demonstrates that the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality. For **PRO-PMs and composite performance measures**, validity should be demonstrated for the computed performance score.

2b3. Exclusions are supported by the clinical evidence; otherwise, they are supported by evidence of sufficient frequency of occurrence so that results are distorted without the exclusion; [12]
AND
If patient preference (e.g., informed decisionmaking) is a basis for exclusion, there must be evidence that the exclusion impacts performance on the measure; in such cases, the measure must be specified so that the information about patient preference and the effect on the measure is transparent (e.g., numerator category computed separately, denominator exclusion category computed separately). [13]

2b4. For outcome measures and other measures when indicated (e.g., resource use):
- **an evidence-based risk-adjustment strategy** (e.g., risk models, risk stratification) is specified; is based on patient factors (including clinical and sociodemographic factors) that influence the measured outcome and are present at start of care; [14] and has demonstrated adequate discrimination and calibration

OR
- rationale/data support no risk adjustment/ stratification.

2b5. Data analysis of computed measure scores demonstrates that methods for scoring and analysis of the specified measure allow for **identification of statistically significant and practically/clinically meaningful** [15] **differences in performance**;
OR
there is evidence of overall less-than-optimal performance.

2b6. If multiple data sources/methods are specified, there is demonstration they produce comparable results.

2b7. For **eMeasures, composites, and PRO-PMs** (or other measures susceptible to missing data), analyses identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias.

2d. For composite performance measures, empirical analyses support the composite construction approach and demonstrate that:
2d1. the component measures fit the quality construct and add value to the overall composite while achieving the related objective of parsimony to the extent possible; and
2d2.the aggregation and weighting rules are consistent with the quality construct and rationale while achieving the related objective of simplicity to the extent possible.
(*if not conducted or results not adequate, justification must be submitted and accepted)*

Notes
10. Reliability testing applies to both the data elements and computed measure score. Examples of reliability testing for data elements include, but are not limited to: inter-rater/abstractor or intra-rater/abstractor studies; internal consistency for multi-item scales; test-retest for survey items. Reliability testing of the measure score addresses precision of measurement (e.g., signal-to-noise).
11. Validity testing applies to both the data elements and computed measure score. Validity testing of data elements typically analyzes agreement with another authoritative source of the same information. Examples of validity testing of the measure score include, but are not limited to: testing hypotheses that the measures scores indicate quality of care, e.g., measure scores are different for groups known to have differences in quality assessed by another valid quality measure or method; correlation of measure scores with another valid indicator of quality for the specific topic; or relationship to conceptually related measures (e.g., scores on process measures to scores on outcome measures). Face validity of the measure score as a quality indicator may be adequate if accomplished through a systematic and transparent process, by identified experts, and explicitly addresses whether performance scores resulting from the measure as specified can be used to distinguish good from poor quality.
12. Examples of evidence that an exclusion distorts measure results include, but are not limited to: frequency of occurrence, variability of exclusions across providers, and sensitivity analyses with and without the exclusion.
13. Patient preference is not a clinical exception to eligibility and can be influenced by provider interventions.
14. Risk factors that influence outcomes should not be specified as exclusions.
15. With large enough sample sizes, small differences that are statistically significant may or may not be practically or clinically meaningful. The substantive question may be, for example, whether a statistically significant difference of one percentage point in the percentage of patients who received smoking cessation counseling (e.g., 74 percent v. 75 percent) is clinically meaningful; or whether a statistically significant difference of $25 in cost for an episode of care (e.g., $5,000 v. $5,025) is practically meaningful. Measures with overall less-than-optimal performance may not demonstrate much variability across providers.

1. DATA/SAMPLE USED FOR ALL TESTING OF THIS MEASURE
Often the same data are used for all aspects of measure testing. In an effort to eliminate duplication, the first five questions apply to all measure testing. If there are differences by aspect of testing,(e.g., reliability vs. validity) be sure to indicate the specific differences in question 1.7.

1.1. What type of data was used for testing? (*Check all the sources of data identified in the measure specifications and data used for testing the measure. Testing must be provided for all the sources of data specified and intended for measure implementation. If different data sources are used for different components in the composite, indicate the component after the checkbox.*)

Measure Specified to Use Data From: (*must be consistent with data sources entered in S.23*)	**Measure Tested with Data From:**
☐ abstracted from paper record	☐ abstracted from paper record
☒ administrative claims	☒ administrative claims
☐ clinical database/registry	☐ clinical database/registry
☐ abstracted from electronic health record	☐ abstracted from electronic health record
☐ eMeasure (HQMF) implemented in EHRs	☐ eMeasure (HQMF) implemented in EHRs
☒ other: Survey – Health care institutions self-report structural characteristics	☒ other: Survey – Health care institutions self-report structural characteristics

1.2. If an existing dataset was used, identify the specific dataset (*the dataset used for testing must be consistent with the measure specifications for target population and healthcare entities being measured; e.g., Medicare Part A claims, Medicaid claims, other commercial insurance, nursing home MDS, home health OASIS, clinical registry*).

15 select states of the Medicaid Extract data (MAX), New York State Medicaid data, American Hospital Association Data, New York State Registry of Regional Perinatal Centers.

1.3. What are the dates of the data used in testing? 2009 - 2011

1.4. What levels of analysis were tested? (*testing must be provided for all the levels specified and intended for measure implementation, e.g., individual clinician, hospital, health plan*)

Measure Specified to Measure Performance of: (*must be consistent with levels entered in item S.26*)	**Measure Tested at Level of:**
☐ individual clinician	☐ individual clinician
☐ group/practice	☐ group/practice
☐ hospital/facility/agency	☐ hospital/facility/agency
☒ health plan	☒ health plan
☒ other: Integrated Delivery System, State, regional, county, national	☒ other: Integrated Delivery System, State, regional, county, national

1.5. How many and which measured entities were included in the testing and analysis (by level of analysis and data source)? (*identify the number and descriptive characteristics of measured entities included in the analysis (e.g., size, location, type); if a sample was used, describe how entities were selected for inclusion in the sample*)

We used MAX data from 15 states abstracted from the Medicaid Analytical Extract (MAX) data set that we prepared and assessed for the PQMP Centers by RTI international. The completeness of data was assessed by the CHOP CoE, who recommended use of 18 states. We were able to acquire hospital data for 15 of those 18 states and they were our population for some testing. Other testing used NY State Medicaid data. The number of plans varied state to state.

1.6. How many and which patients were included in the testing and analysis (by level of analysis and data source)? (*identify the number and descriptive characteristics of patients included in the analysis (e.g., age, sex, race, diagnosis); if a sample was used, describe how patients were selected for inclusion in the sample*)

This was an analysis of women who gave birth (MAX data) and women who gave birth and newborns of both sexes using the NY State data.

In New York State, 48.6% of deliveries in 2011 occurred in women insured by Medicaid (16). In our analysis year, 55.6% (4197 neonates) of low birthweight neonates admitted to NICUs across New York State and who were in our study of newborn temperatures (approximately 90% of all newborns admitted to level 2 or 3 nurseries) were insured by Medicaid.

1.7. If there are differences in the data or sample used for different aspects of testing (e.g., reliability, validity, exclusions, risk adjustment), identify how the data or sample are different for each aspect of testing reported below.

MAX data could only be used for the OB construct and the NICU construct and only among Class A eligible. We are attempting to re-run the data to combine the two constructs but were unable to have it in time for this submission. We are also seeking to have the NYS DoH re-run data that will allow for simultaneous assessment of three of the four components but they have not yet been able to do so and we are not certain that they will be able to in time for the meeting, but we are hopeful.

1.8 What were the patient-level sociodemographic (SDS) variables that were available and analyzed in the data or sample used? For example: patient-reported data (e.g., income, education, language), proxy variables when SDS data are not collected from each patient (e.g. census tract), or patient community characteristics (e.g. percent vacant housing, crime rate).

Race/ethnicity, county of residence was used to develop optional urbanicity/rurality, and poverty variables as described in the specifications.

_____

2a2. RELIABILITY TESTING

2a2.1. What level of reliability testing was conducted?

Note*: Current guidance for composite measure evaluation states that reliability must be demonstrated for the composite performance measure score.*

☒ **Performance measure score** (e.g., *signal-to-noise analysis*)

2a2.2. Describe the method of reliability testing and what it tests (*describe the steps—do not just name a method; what type of error does it test; what statistical analysis was used*)

The strengths of this measure derive from its systematic development, its meticulous specification, its careful conceptualization and articulation and its grounding in existing science and consensus.

The data collection and reliability therein depend upon the use of administrative data. These data are used to identify deliveries (our specifications are a slight enhancement of CDC methodologies described in Kuklina et. al (38); to Kuklina's work we added Revenue code 722). The CDC is using this method as its standard for identifying deliveries from administrative data.

This was important for our test because the Medicaid MAX data provided by CMS and in which these schemas were tested does not include DRGs, which are employed in the Kuklina method. We tested also a variation of the approach to identify deliveries employed by HEDIS in its Timing of Prenatal Care measure in the initial CHIPRA core set. We found that these approaches identified substantially the same population of deliveries in a fifteen state subset of the national MAX database. We chose the 15 states to include in an attempt to manifest some standardization of approaches across the seven AHRQ-CMS CHIPRA Centers of Excellence—they were recommended to us as a diverse set of states with high data quality by the Children's Hospital of Pennsylvania Center which has used them extensively in a number of their validation activities. As the different approaches produced 90% or more overlap, we decided to specify the measure based upon the Kuklina/CDC approach as it is both widely-used and relevant for the type of population-based approach to measurement proposed in this measure. The method improved the previous gold standard for identifying deliveries by identifying an additional 958,000 deliveries, beyond the previously identified 27,128,539. True estimates of Sensitivity and specificity are not available as there is no other gold standard.

In determining which women were to be considered potentially in need of HROB services, our specifications further rely upon administrative data. One study found that quality measures that could be calculated using administrative data showed higher rates of performance than indicated by a review of the medical record alone, and that claims data is more accurate for identifying services with a high likelihood of documentation due to reimbursement (39). While such findings are not specific to an obstetrical population the issue at hand is actually the identification fo the chronic diseases Further, at the current stage of EMR development and implementation, chart review is likely to prove infeasible for population-based measures of this scope. Since this measure is specified to be interpreted at the population and not the individual level, the impact of some of the imperfections of using administrative data will be overcome naturally because of the law of large numbers.

As an illustration of our approach, we provide a case example of our decision to exclude two diagnoses from the inclusion criteria. The Expert Panel rated valvular heart disease as significant and an indication of the need for HROB. In its deliberations it made clear that often trivial conditions such as murmurs or simple mitral valve prolapse were not the target of its rating. So in specifying the inclusion specifications, we included Clinical Classifications category 96, Heart

Valve Disorders, but specified the removal of three ICD-9 codes from that category (4240 Mitral Valve Disorders, 7852 Undiagnosed Cardiac Murmurs, and 7853 Other Abnormal Heart Sounds). The CAPQuaM team made the clinical judgment that it was more true to the intention of the panel to accept the error that results from eliminating the rarer more serious isolated mitral valve disorders than the error of including the common and often innocuous mitral valve prolapse in the specified sample. In addition, we want to be clear that the three ICD-9 codes mentioned were not then used as exclusion criteria if there were other reasons for the pregnancy to be identified as high risk. Rather these codes were removed from the inclusion criteria.

Regarding the assessment of the presence or absence of structural characteristics in this measure set, we have specified this measure to use the results of questionnaires or surveys that we envision as paper, email or internet-based. Our feasibility assessment determined that these data are readily available from key individuals at the hospitals. We imagine that one or more states or health plans have databases that link some or all of these data (especially Level 3 or higher nurseries). The regular use of Revenue Code 173 or 174 could also be used to identify Level 3 or 4 Nursery care, respectively. If challenged, we consider public self-report to be preferable to the use of a database unless there is evidence of deception or fraud.

We have developed our survey questions in accordance with best practices and after studying the American Hospital Association Annual Survey of Hospitals (17, 40), which is considered the authoritative survey of hospital structural characteristics in the United States. After careful internal review and revision by the scientific team, appropriate clinicians, and experts, we concluded our development of the four-item questionnaire with one formal cognitive interview. We conducted this interview with the Director of Special Projects in the Office of Patient Excellence at the Mount Sinai Medical Center. This individual is not a clinician and had no previous involvement with the development of these measures. The items were revised and modified in accordance with the findings from that interview and provided back for her review. After a second round of revisions we received confirmation that all relevant issues had been addressed successfully and that the questions were clear and unambiguous. Further review by individuals not associated with the team confirmed this feedback. The questions are modeled after similar standard questions that have been used for years by the American Hospital Association.

2a2.3. What were the statistical results from reliability testing? (e.*g., percent agreement and kappa for the critical data elements; distribution of reliability statistics from a signal-to-noise analysis*)
Confidence intervals were very narrow as illustrated by:

Numerator is Level 3 Obstetrical services:

State	N Obs	N	Mean	Std Dev	Lower 95% CL for Mean	Upper 95% CL for Mean
AZ	**51527**	51141	0.33816	0.47309	0.33406	0.34226
ID	**9553**	9003	0.19427	0.39566	0.18609	0.20244
IL	**60073**	17164	0.24715	0.43136	0.24069	0.2536
IN	**36656**	36433	0.25131	0.43377	0.24686	0.25576
KS	**14410**	13678	0.52881	0.49919	0.52044	0.53717
KY	**28726**	28708	0.11613	0.32039	0.11243	0.11984
LA	**37731**	1376	0.43314	0.49569	0.40693	0.45935
MO	**33287**	27404	0.15863	0.36533	0.1543	0.16295
MT	**4016**	3997	0.35352	0.47812	0.33869	0.36834
NC	**65324**	65321	0.47032	0.49912	0.4665	0.47415

NH	**3792**	3650	0.12411	0.32975	0.11341	0.13481
NJ	**14363**	13111	0.32797	0.46949	0.31993	0.33601
NM	**14849**	14038	0.31842	0.46588	0.31071	0.32613
NY	**116151**	96499	0.16376	0.37006	0.16143	0.1661
VA	**31355**	30701	0.35015	0.47703	0.34482	0.35549
VT	**2612**	2361	0.37315	0.48374	0.35362	0.39267
WI	**23796**	18965	0.14854	0.35564	0.14347	0.1536
WY	**3242**	3239	0.046	0.20952	0.03878	0.05322

Results for Level 3 NICU are similar with all but 2 states having standard errors less than 1%.

We found that of about 119,000 Medicaid deliveries in New York State in 2010, 59,254 were at sufficiently elevated risk to qualify for this measure set: 56,465 (~47%) were identified using Class A criteria, 7,800 using Class B (~7%), and 59,254 (just under 50%) using either Class A or Class B, meaning that 2,789 (or about 5% of the overall high-risk pregnancies) were identified only using Class B. We expected a substantial "voltage drop" between a condition of elevated risk and a complication or an undesirable outcome. Hence maternal diagnoses codes of Class A will predictably be orders of magnitude larger than the delivery and neonatal codes of Class B. These findings are consistent with our predictions and expectations. Our team had predicted that 40-50% of all pregnancies would have elevated risk and these findings are consistent with the expectations that Medicaid would be at least at the higher end of that range. Use of a mother-only algorithm in MAX data in 15 states indicates the proportion of high-risk pregnancies ranges from 31.50% in NJ to 63.97% in KY. The NY MAX finding was 55,379 HROB pregnancies, almost identical to the 56,465 found using internal data bases on the maternal codes, indicating very high reliability across systems. This is reliability at the level of estimation of the denominator size.

A few examples of data element assessments are cited below:
(1) Lois Quam, Lynda B. M. Ellis, Pat Venus, Jon Clouse, Cynthia G. Taylor and Sheila Leatherman. Using Claims Data for Epidemiologic Research: The Concordance of Claims-Based Criteria with the Medical Record and Patient Survey for Identifying a Hypertensive Population. Medical Care. Vol. 31, No. 6 (Jun., 1993), pp. 498-507. A sample of 2,079 patients from two study sites with medical service or pharmacy claims indicating a diagnosis of essential hypertension were surveyed, and the medical records of 182 of the 1,275 survey respondents were reviewed. Using medical record data was 74 % agreement with either the medical record or the patient survey.

(2) Machelle Wilchesky, Robyn M Tamblyn, Allen Huang. Validation of diagnostic codes within medical services claims. Journal of Clinical Epidemiology. February 2004. Volume 57, Issue 2, Pages 131–141.

The goal of this study is to determine the sensitivity and specificity of medical services claims diagnoses for surveillance of 14 drug disease contraindications used in drug utilization review, the Charlson comorbidity index and the Johns Hopkins Adjusted Care Group Case-Mix profile (ADGs). Diagnoses were abstracted from the medical charts of 14,980 patients, and were used as the "gold standard," against which diagnoses obtained from the administrative database for the same patients were compared. The administrative data was found to have diagnoses and conditions that were highly specific but that varied by condition in terms of sensitivity. To appropriately obtain diagnostic profiles, it is recommended that data pertaining to all physician billings be used, as we do for this measure.

Fowles, Jinnet B. PhD; Fowler, Elizabeth J. PhD; Craft, Cheryl RN. Validation of Claims Diagnoses and Self-Reported Conditions Compared with Medical Records for Selected Chronic Diseases. Journal of Ambulatory Care Management: January 1998 - Volume 21 - Issue 1. his article assesses the validity of ambulatory administrative and encounter data

and patient self-reported information compared with information contained in the ambulatory medical record for 17 chronic diseases. Sensitivity and specificity for claims and self-report were calculated for each chronic condition using the medical record as the criterion standard. The analysis was performed first by blinded review and then repeated with an unblinded review. Across 17 studied conditions, specificity was consistently greater than 0.95, indicating that we are unlikely to falsely identify women as being of high risk.

A Comparison of Ambulatory Medicaid Claims to Medical Records: A Reliability Assessment. American Journal of Medical Quality June 1998 13: 63-69. This study compares the documentation of ambulatory care visits and diagnoses in Medicaid paid claims and in medical records… Medicaid are accurate and useful for examining average ambulatory use patterns, but may be subject to bias when used for clinical profiles at the level of physicians.

2a2.4 What is your interpretation of the results in terms of demonstrating reliability? (i.*e., what do the results mean and what are the norms for the test conducted?*)

We conclude that the measures are reliable and feasible with variations as expected. We also conclude that since 95% of cases are identifiable with Class A criteria, that the inability to link maternal and infant charts does not constitute a rationale for exclusion from this measure.

_____

2b2. VALIDITY TESTING

Note*: Current guidance for composite measure evaluation states that validity should be demonstrated for the composite performance measure score. If not feasible for initial endorsement, acceptable alternatives include assessment of content or face validity of the composite OR demonstration of validity for each component. Empirical validity testing of the composite measure score is expected by the time of endorsement maintenance.*

2b2.1. What level of validity testing was conducted?

☐ **Composite performance measure score**

☒ **Empirical validity testing**

☒ **Systematic assessment of face validity of performance measure score as an indicator** of quality or resource use (*i.e., is an accurate reflection of performance on quality or resource use and can distinguish good from poor performance*)

☒ **Systematic assessment of content validity**

☐ **Validity testing for component measures** *(check all that apply)*

Note*: applies to ALL component measures, unless already endorsed or are being submitted for individual endorsement.*

☐ **Endorsed (or submitted) as individual performance measures**

☒ **Critical data elements** (*data element validity must address ALL critical data elements*)

☐ **Empirical validity testing of the component measure score(s)**

☒ **Systematic assessment of face validity of component measure score(s) as an indicator** of quality or resource use (*i.e., is an accurate reflection of performance on quality or resource use and can distinguish good from poor performance*)

2b2.2. For each level of testing checked above, describe the method of validity testing and what it tests (*describe the steps—do not just name a method; what was tested, e.g., accuracy of data elements compared to authoritative source, relationship to another measure as expected; what statistical analysis was used)*

The reliability section above also contains information related to validity.

Our definition of high-risk obstetrical services results from a formal RAND/UCLA modified Delphi process conducted with a multidisciplinary panel of national experts that included obstetricians, MFM specialists, and a nurse midwife,

anesthesiologist and family physician. We carefully operationalized the panel's clinical recommendations by fine tuning AHRQ's Clinical Classification Software. We operationalized panel specifications using data elements that are available in typical administrative data sets. Potential exceptions are elements like race and ethnicity. Chronic diseases are well represented in claims data and generally well captured, especially over time.

Use of administrative data in performance assessment is common. They contain consistent elements, are available, inform regarding large numbers of individuals, and are relatively inexpensive. Validity of many has been established, and their strengths and weaknesses relative to data abstracted from medical records and obtained via survey have been documented and their use encouraged by federal agencies (41). The Centers for Medicare & Medicaid Services has made clear to the participating AHRQ-CMS CHIPRA Centers of Excellence funded to develop measures in the Pediatric Quality Measures Program that it places a premium on feasibility. Expert Panels have been demonstrated to enhance measure development and health care evaluation, including for children (42); frontline practitioners can assist researchers to create useful measures (43). CAPQuaM's 360 degree method is highly engaged with collaborators, partners, and the literature. It targets relevant information and perspective and measures emerge from the process.

At a performance level, the measures were predicted and found to vary with urbanicity (higher levels of availability) and rurality (lower levels of availability) reflecting the distribution of services across rural and urban areas. It is normative for services to be less available in rural than urban areas and our measures confirmed this. We conclude face validity from our observations.

State to state performance varied as expected given our knowledge of rurality and distribution of clinical services.

Key aspects of validity of HROB measures

Availability

The construct of availability is complex and can be muddied in the distinction or lack thereof between availability, access, and utilization (27). For this first PQMP measure set on availability of HROB services, we selected four constructs that avoid any potential confusion between availability and access or utilization. In modern medical practice, all women having babies require some form of delivery services. By looking at the rate at which eligible (i.e., high-risk) deliveries that occur in hospitals with key structural elements are associated with better outcomes, we created an index of the availability of those services. All else equal, we would expect women who live in more medically dense communities to experience greater availability than those who live in less medically dense communities and those who live in more isolated communities to have even less availability. Within geographic region, these measures are sensitive to the contractual, travel, and other arrangements that a system makes aviable for its population of pregnant women.

High-risk

We have operationalized a systematic expert process informed by a detailed literature review and incorporating a well-described and frequently-utilized system developed by AHRQ. While we have modified this system, it has been done to be consistent with its use in this context and to remain consistent with the guidance of the Expert Panel. It is transparent and has high face validity.

Our definition that the need for high-risk services extends from preconception to the puerperium implies that maternal conditions (comorbidities), complications of pregnancy, and complications of delivery each may be used to identify the need for high-risk services. For these measures we classify risk in two ways, one based upon maternal

diagnoses and another based upon delivery, fetal, or infant conditions. We note that all low birthweight infants are products of a high-risk pregnancy, since premature labor and growth retardation are within our definition of risk. While linking infant and mother charts may occasionally be a challenge for hospitals, it should be less challenging for reporting entities. Our work with the New York State Medicaid data has confirmed the feasibility of such linkage. If linkage is not possible, the Class A portion of the measure can be calculated based upon maternal records alone.

Structural Aspects of Care

Data regarding structural aspects are self-reports from health care facilities. We developed a 4-item questionnaire with internal review and a single cognitive interview with follow-up review by an individual who could be called upon to complete such a survey at Mount Sinai. Medical Center. Our questions are specific and factual. Self-report is the current standard for assessing facility characteristics, frequently through the use of the American Hospital Association Survey cited above. The lack of anonymity for the person completing the survey and the potential verifiability of the questions enhance validity.

We have cited abundant literature and guidelines that consistently hold that the structural aspects that are the targets of these measures matter in terms of maternal and infant outcomes. In data from New York State Medicaid among women who met our criteria for high-risk deliveries, we found that these measures vary with a gradient of accessibility of medical services as associated with geographic proximity or metropolitan areas. See Table 6 below.

2b2.3. What were the statistical results from validity testing? (*e.g., correlation; t-test*)

TABLE 6

HROB Summary (Combined Unduplicated) New York State Medicaid, 2010					
Urbanicity	UIC	N	OB Proxy[1]	Transfusion Proxy[2]	NICU >=3[3]
URBAN					
Large Metropolitan	1	48,562	27.10%	14.62%	37.98%
Small Metropolitan	2	7229	15.95%	10.26%	21.69%
RURAL					
Adjacent to Large Metro	3,4	796	12.64%	6.67%	7.79%
Adjacent to Small Metro	5,6 7	2001	3.31%	3.45%	7.30%
Not Adjacent	8,9	616	1.14%	1.30%	1.46%

1. Proxy for OB 24/7 coverage is Level 3 Obstetrical Care according to the American Hospital Association Survey.
2. Proxy for 24/7 Tranfusion sevices is designation s a Regional Perinatal center by New York State, for which 24 /7 transfusion services are a requirement
3. NICU represents provision of Level 3 or Level 4 NICU services as identified by frequent the billing of Medicaid for Level 3 or Level 4 NICU services using Revenue Codes 173 or 174.

Recent data updates these findings. CAPQuaM is working with Drs. Jill Mhyre (an Expert Panel member), Andrea Fuller, and Brenda Bucklin on a manuscript, "Anesthesia Services for High-risk Obstetrics: Results from the 2011 Obstetric Anesthesia Workforce Survey." This survey supports the salience of this measure. Results are shown in Tables 2-4 below. The data within categories are nationally representative, but final sampling weights are not ready to make national estimates across categories.

TABLE 2

	High-risk referral center	Not a high-risk referral	P-value
Facility Type			
Community hospital	61(58.7%)	207 (91.6%)	<0.001
Anesthesia residency	42 (40.4%)	6 (2.7%)	
CRNA training program	1 (1%)	4 (1.8%)	
Military	0	2 (0.9%)	
Rural critical access	0	7 (3.1%)	
Annual Delivery volume*	3000 [2000, 4500]	750 [350,1800]	<0.001
Cesarean Deliver y rate *	30% [27, 35]	29% [25, 33]	0.01

Proportion funded by private insurance*	45% [25, 65]	35% [20, 70]	0.22

TABLE 3

		High-risk referral	Not a high-risk	P-value
Night-time and weekend staffing				
	In-house anesthesiologist or anesthesia team dedicated to L&D	61(60.4%)	57 (26.3%)	<0.001
	In-house anesthesiologist or anesthesia	25 (24.8%)	27 (12.4%)	
	In-house independently-practicing CRNA	6 (5.9%)	5 (2.3%)	
	In-house independently-practicing CRNA	0	6 (2.8%)	
	Anesthesiologist or CRNA on call from home	9 (8.9%)	12 2 (56.2%)	
Antepartum consultation service		89 (89.0%)	140 (65.4%)	<0.001
Massive transfusion protocol		86 (95.6%)	113 (56.0%)	<0.001

Variability in anesthesiology staffing is not defined by hospital characteristics. While similar structural characteristics predict obstetrical and anesthesiology coverage they do not overlap, supporting distinct measures for OB and anesthesiology coverage.

TABLE 4

	Self-reported characteristics of the respondent's institution				
	High-risk referral center ≥1500 births per year	High-risk referral center <1500 births per year	Low risk, ≥1500 births per year	Low risk, <1500 births per year	*P-value*
Recommended anesthesia staffing					
In-house anesthesiologist	55 (68.8%)	3 (30%)	26 (44.8%)	19 (13.7%)	<0.001
In-house medically directed resident or CRNA	19 (23.8%)	2 (20%)	16 (27.6%)	15 (10.8%)	
In-house CRNA without medical direction	4 (5%)	4 (40%)	5 (8.6%)	20 (14.4%)	
In-hospital coverage unnecessary	2 (2.5%)	1 (10%)	11 (19.0%)	85 (61.2%)	
Recommended obstetric staffing					
In-house obstetrician	75 (93.8%)	8 (80%)	41 (69.5%)	43 (30.9%)	0.001
In-house nurse midwife	0	0	1 (1.6%)	5 (3.6%)	
In-house other physician	0	1 (10%)	0	2 (1.4%)	
In-hospital coverage unnecessary	5 (6.3%)	1 (10%)	17 (28.8%)	89 (64.0%)	

Our transfusion measure incorporates language from the NY State DOH criteria to identify Regional Perinatal Centers in NY. The clinical imperative to look at availability of these services is set forth by the California Maternal Quality Care Coalition (CMQCC) (13). Hemorrhages occur predictably in the context of coagulation disorder, somewhat predictably when problems of placentation may be noted before or early in labor, or unpredictably. Large amounts of blood loss may go unnoticed or unappreciated if not monitored, sought, and understood by experienced and meticulous clinicians, often aided by thoughtful protocols. And even in the hands of excellent clinicians, the management of hemorrhage requires early recognition, proper management to achieve rapid hemostasis, and prompt and sometimes repeated transfusion. Key data from CMQCC are shown in Table 5 below.

TABLE 5

Frequency of Delays in obtaining blood products when needed By Individual Respondents n (% by row)

Hospital Size: # Live Births (2005)	No Delay	No Blood Bank on-site	Lack of Pre-natal Record	Blood Bank hesitancy to release O-	Blood bank closed/off hours	Blood Bank is busy	Total
<1000 (33)	24 (73)	2 (6)	0 (0)	2 (6)	0 (0)	3 (9)	31 (94)
1001-3000 (121)	59 (49)	6 (5)	9 (7)	15 (12)	0 (0)	13 (11)	102 (84)
>3000 (86)	47 (55)	2 (2)	5 (6)	13 (15)	0 (0)	12 (14)	79 (92)
Total (240)	130 (54)	10 (4)	14 (6)	30 (12)	0 (0)	28 (12)	212* (88)

2b2.4. What is your interpretation of the results in terms of demonstrating validity? (i.*e., what do the results mean and what are the norms for the test conducted?*)

We interpret the findings to suggest that these services become less available with increasing rurality, as we had predicted. We designed the measures to identify reduced availability for any reason (including geographic isolation) and the observed gradient strongly supports the validity of these as population measures of availability. These components move in similar directions but not in lock step, confirming that they are measuring related but not identical constructs, as we would hope. The overall availability of these structural components of high-risk obstetrical services is low compared to the identified need. National work force data presented confirms the critical importance of the anesthesiologist construct that was not included in the NY State test.

_____

2b3. EXCLUSIONS ANALYSIS

Note: *Applies to the composite performance measure, as well all component measures unless they are already endorsed or are being submitted for individual endorsement.*

NA ☐ no exclusions — *skip to section 2b4*

2b3.1. Describe the method of testing exclusions and what it tests (*describe the steps—do not just name a method; what was tested, e.g., whether exclusions affect overall performance scores; what statistical analysis was used*)

We compared the number of eligible births in NY State with Class A and Class B to those with Class A only and conclude there is no need to exclude if maternal and infant charts cannot be linked.

2b3.2. What were the statistical results from testing exclusions? (*include overall number and percentage of individuals excluded, frequency distribution of exclusions across measured entities, and impact on performance measure scores*)

2b3.3. What is your interpretation of the results in terms of demonstrating that exclusions are needed to prevent unfair distortion of performance results? (*i.e., the value outweighs the burden of increased data collection and analysis. Note: **If patient preference is an exclusion**, the measure must be specified so that the effect on the performance score is transparent, e.g., scores with and without exclusion*)

2b4. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES

Note: *Applies to all outcome or resource use component measures, unless already endorsed or are being submitted for individual endorsement.*

If not an intermediate or health outcome, or PRO-PM, or resource use measure, skip to section 2b5.

2b4.1. What method of controlling for differences in case mix is used? *(check all that apply)*

☐ **Endorsed (or submitted) as individual performance measures**

☐ **No risk adjustment or stratification**

☐ **Statistical risk model**

☐ **Stratification by risk categories**

☒ **Other,** We call for stratification by race/ethnicity as called for in the CHIPRA legislation that funded our work. We further specify for optional stratification by Poverty and Rurality that we recommend but do not require. These are not for purposes of risk adjustment but allow for more granular understanding of the data and the detection of racial and ethnic disparities.

2b4.2. If an outcome or resource use component measure is not risk adjusted or stratified, provide rationale and analyses to demonstrate that controlling for differences in patient characteristics (case mix) is not needed to achieve fair comparisons across measured entities.

This is a population index of availability of care. The structural characteristics included in this measure should be uniformly available across risk adjustment categories. We call for stratification to produce additional information rather than to risk adjust.

2b4.3. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors used in the statistical risk model or for stratification by risk (*e.g., potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of p<0.10; correlation of x or higher; patient factors should be present at the start of care*)

2b4.4a. What were the statistical results of the analyses used to select risk factors?

2b4.4b. Describe the analyses and interpretation resulting in the decision to select SDS factors (e.g. prevalence of the factor across measured entities, empirical association with the outcome, contribution of unique variation in the outcome, assessment of between-unit effects and within-unit effects)

2b4.5. Describe the method of testing/analysis used to develop and validate the adequacy of the statistical model or stratification approach (*describe the steps—do not just name a method; what statistical analysis was used*)

Provide the statistical results from testing the approach to controlling for differences in patient characteristics (case mix) below.
if stratified, skip to 2b4.9

2b4.6. Statistical Risk Model Discrimination Statistics (*e.g., c-statistic, R-squared*)**:**

2b4.7. Statistical Risk Model Calibration Statistics (*e.g., Hosmer-Lemeshow statistic*):

2b4.8. Statistical Risk Model Calibration – Risk decile plots or calibration curves:

2b4.9. Results of Risk Stratification Analysis:

2b4.10. What is your interpretation of the results in terms of demonstrating adequacy of controlling for differences in patient characteristics (case mix)? (i.*e., what do the results mean and what are the norms for the test conducted?*)

***2b4.11. Optional Additional Testing for Risk Adjustment** (*not required, but would provide additional support of adequacy of risk model, e.g., testing of risk model in another data set; sensitivity analysis for missing data; other methods that were assessed*)

2b5. IDENTIFICATION OF STATISTICALLY SIGNIFICANT & MEANINGFUL DIFFERENCES IN PERFORMANCE
Note: *Applies to the composite performance measure.*

2b5.1. Describe the method for determining if statistically significant and clinically/practically meaningful differences in performance measure scores among the measured entities can be identified (*describe the steps—do not just name a method; what statistical analysis was used? Do not just repeat the information provided related to performance gap in 1b)*

Analysis of variance using generalized models and SAS software demonstrated state to state differences, racial and ethnic differences, and differences between rural, urban, and suburban areas for the constructs that make up this measure.

Charts illustrating state to state differences in two of the components are shown above.

2b5.2. What were the statistical results from testing the ability to identify statistically significant and/or clinically/practically meaningful differences in performance measure scores across measured entities? (e.g., *number and percentage of entities with scores that were statistically significantly different from mean or some benchmark, different from expected; how was meaningful difference defined*)

2b5.3. What is your interpretation of the results in terms of demonstrating the ability to identify statistically significant and/or clinically/practically meaningful differences in performance across measured entities? (i.*e., what do the results mean in terms of statistical and meaningful differences?*)

The measure is sufficiently sensitive to identify meaningful differences in availability.

2b6. COMPARABILITY OF PERFORMANCE SCORES WHEN MORE THAN ONE SET OF SPECIFICATIONS
Note: *Applies to all component measures, unless already endorsed or are being submitted for individual endorsement.*
If only one set of specifications for each component, this section can be skipped.

Note: *This item is directed to measures that are risk-adjusted (with or without SDS factors)* ***OR*** *to measures with more than one set of specifications/instructions (e.g., one set of specifications for how to identify and compute the measure from medical record abstraction and a different set of specifications for claims or eMeasures). It does not apply to measures that use more than one source of data in one set of specifications/instructions (e.g., claims data to identify the denominator and medical record abstraction for the numerator).* ***Comparability is not required when comparing performance scores with and without SDS factors in the risk adjustment model. However, if comparability is not demonstrated for measures with more than one set of specifications/instructions, the different specifications (e.g., for medical records vs. claims) should be submitted as separate measures.***

2b6.1. Describe the method of testing conducted to demonstrate comparability of performance scores for the same entities across the different data sources/specifications (*describe the steps—do not just name a method; what statistical analysis was used*)

2b6.2. What were the statistical results from testing comparability of performance scores for the same entities when using different data sources/specifications? (*e.g., correlation, rank order*)

2b6.3. What is your interpretation of the results in terms of demonstrating comparability of performance measure scores for the same entities across the different data sources/specifications? (i.*e., what do the results mean and what are the norms for the test conducted?*)

2b7. MISSING DATA ANALYSIS AND MINIMIZING BIAS

Note: *Applies to the overall composite measure.*

2b7.1. Describe the method of testing conducted to identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias (*describe the steps—do not just name a method; what statistical analysis was used*)

2b7.2. What is the overall frequency of missing data, the distribution of missing data across providers, and the results from testing related to missing data? (*e.g., results of sensitivity analysis of the effect of various rules for missing data/nonresponse; if no empirical sensitivity analysis, identify the approaches for handling missing data that were considered and pros and cons of each*)

2b7.3. What is your interpretation of the results in terms of demonstrating that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias**?** (i.*e., what do the results mean in terms of supporting the selected approach for missing data and what are the norms for the test conducted; if no empirical analysis, provide rationale for the selected approach for missing data*)

2d. EMPIRICAL ANALYSIS TO SUPPORT COMPOSITE CONSTRUCTION APPROACH

Note: *If empirical analyses do not provide adequate results—or are not conducted—justification must be provided and accepted in order to meet the must-pass criterion of Scientific Acceptability of Measure Properties. Each of the following questions has instructions if there is no empirical analysis.*

2d1. Empirical analysis demonstrating that the component measures fit the quality construct, add value to the overall composite, and achieve the object of parsimony to the extent possible.

2d1.1 Describe the method used (*describe the steps—do not just name a method; what statistical analysis was used; if no empirical analysis, provide justification)*

2d1.2. What were the statistical results obtained from the analysis of the components? (e.g., *correlations, contribution of each component to the composite score, etc*.; *if no empirical analysis, identify the components that were considered and the pros and cons of each*)

2d1.3. What is your interpretation of the results in terms of demonstrating that the components included in the composite are consistent with the described quality construct and add value to the overall composite? (i.*e., what do the results mean in terms of supporting inclusion of the components; if no empirical analysis, provide rationale for the components that were selected)*

2d2. Empirical analysis demonstrating that the aggregations and weighting rules are consistent with the quality construct and achieve the objective of simplicity to the extent possible

2d2.1 Describe the method used (*describe the steps—do not just name a method; what statistical analysis was used; if no empirical analysis, provide justification)*

2d2.2. What were the statistical results obtained from the analysis of the aggregation and weighting rules? (e.g., *results of sensitivity analysis of effect of different aggregations and/or weighting rules; if no empirical analysis, identify the aggregation and weighting rules that were considered and the pros and cons of each*)

2d2.3. What is your interpretation of the results in terms of demonstrating the aggregation and weighting rules are consistent with the described quality construct? (i.*e., what do the results mean in terms of supporting the selected rules for aggregation and weighting; if no empirical analysis, provide rationale for the selected rules for aggregation and weighting*)

3. Feasibility

Extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

3a. Byproduct of Care Processes

For clinical measures, the required data elements are routinely generated and used during care delivery (e.g., blood pressure, lab test, diagnosis, medication order).

3a.1. **Data Elements Generated as Byproduct of Care Processes.**

Coded by someone other than person obtaining original information (e.g., DRG, ICD-9 codes on claims), Other

If other: Survey of the facility

3b. Electronic Sources

The required data elements are available in electronic health records or other electronic sources. If the required data are not in electronic health records or existing electronic sources, a credible, near-term path to electronic collection is specified.

3b.1. **To what extent are the specified data elements available electronically in defined fields?** *(i.e., data elements that are needed to compute the performance measure score are in defined, computer-readable fields)*

Some data elements are in defined fields in electronic sources

3b.2. **If ALL the data elements needed to compute the performance measure score are not from electronic sources, specify a credible, near-term path to electronic capture, OR provide a rationale for using other than electronic sources.**

A registry of structural characteristics of all facilities that provide for birthing and/or delivery could be readily develop functions at hospitals giving birth could be developed.

3b.3. **If this is an eMeasure, provide a summary of the feasibility assessment in an attached file or make available at a measure-specific URL.**

Attachment:

3c. Data Collection Strategy

Demonstration that the data collection strategy (e.g., source, timing, frequency, sampling, patient confidentiality, costs associated with fees/licensing of proprietary measures) can be implemented (e.g., already in operational use, or testing demonstrates that it is ready to put into operational use). For eMeasures, a feasibility assessment addresses the data elements and measure logic and demonstrates the eMeasure can be implemented or feasibility concerns can be adequately addressed.

3c.1. **Describe what you have learned/modified as a result of testing and/or operational use of the measure regarding data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, time and cost of data collection, other feasibility/implementation issues.**

IF a PRO-PM, consider implications for both individuals providing PROM data (patients, service recipients, respondents) and those whose performance is being measured.

The measure was developed in tandem with feasibility assessment and is structured to be feasible and readily captured. We learned that in lieu of the survey measures, it was possible to use proxies which were better for some (NICU) than others (not available for anesthesiologist). A regular survey and registry of these features would make measurement easier in an ongoing fashion rather than periodic collection of the survey as would be necessary today.

3c.2. **Describe any fees, licensing, or other requirements to use any aspect of the measure as specified** *(e.g., value/code set, risk model, programming code, algorithm)*.

None at present.

4. Usability and Use

Extent to which potential audiences (e.g., consumers, purchasers, providers, policy makers) are using or could use performance results for both accountability and performance improvement to achieve the goal of high-quality, efficient healthcare for individuals or populations.

4a. Accountability and Transparency

Performance results are used in at least one accountability application within three years after initial endorsement and are publicly reported within six years after initial endorsement (or the data on performance results are available). If not in use at the time of initial endorsement, then a credible plan for implementation within the specified timeframes is provided.

4.1. Current and Planned Use

NQF-endorsed measures are expected to be used in at least one accountability application within 3 years and publicly reported within 6 years of initial endorsement in addition to performance improvement.

Planned	Current Use (for current use provide URL)
Not in use Use Unknown	

4a.1. For each CURRENT use, checked above, provide:

- Name of program and sponsor
- Purpose
- Geographic area and number and percentage of accountable entities and patients included

4a.2. If not currently publicly reported OR used in at least one other accountability application (e.g., payment program, certification, licensing) what are the reasons? (*e.g., Do policies or actions of the developer/steward or accountable entities restrict access to performance results or impede implementation?*)

The measure is not currently in widespread use because it is new and awaiting endorsement. It is a Pediatric Quality Measures Program Measure and is widely anticipated for future use.

4a.3. If not currently publicly reported OR used in at least one other accountability application, provide a credible plan for implementation within the expected timeframes -- any accountability application within 3 years and publicly reported within 6 years of initial endorsement. (*Credible plan includes the specific program, purpose, intended audience, and timeline for implementing the measure within the specified timeframes. A plan for accountability applications addresses mechanisms for data aggregation and reporting.*)

As a part of our work in the PQMP we are developing specific plans for dissemination and use. Our plan includes submitting for endorsement of the NQF for the measure. We are having conversations with partners regarding the application and use of the measure.

4b. Improvement

Progress toward achieving the goal of high-quality, efficient healthcare for individuals or populations is demonstrated. If not in use for performance improvement at the time of initial endorsement, then a credible rationale describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.

4b.1. Progress on Improvement. (Not required for initial endorsement unless available.)

Performance results on this measure (current and over time) should be provided in 1b.2 and 1b.4. Discuss:

- **Progress (trends in performance results, number and percentage of people receiving high-quality healthcare)**
- **Geographic area and number and percentage of accountable entities and patients included**

4b.2. If no improvement was demonstrated, what are the reasons? If not in use for performance improvement at the time of initial endorsement, provide a credible rationale that describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.

4c. Unintended Consequences

The benefits of the performance measure in facilitating progress toward achieving high-quality, efficient healthcare for individuals or populations outweigh evidence of unintended negative consequences to individuals or populations (if such evidence exists).

4c.1. Were any unintended negative consequences to individuals or populations identified during testing; OR has evidence of unintended negative consequences to individuals or populations been reported since implementation? If so, identify the negative unintended consequences and describe how benefits outweigh them or actions taken to mitigate them.

This measure is a population index and not an individual measure of quality. To date there is no evidence suggesting untoward or unintended consequences of the measure or its use.

5. Comparison to Related or Competing Measures

If a measure meets the above criteria and there are endorsed or new related measures (either the same measure focus or the same target population) or competing measures (both the same measure focus and the same target population), the measures are compared to address harmonization and/or selection of the best measure.

5. Relation to Other NQF-endorsed Measures
Are there related measures (conceptually, either same measure focus or target population) or competing measures (conceptually both the same measure focus and same target population)? If yes, list the NQF # and title of all related and/or competing measures.
No

5.1a. **List of related or competing measures (selected from NQF-endorsed measures)**

5.1b. **If related or competing measures are not NQF endorsed please indicate measure title and steward.**

5a. Harmonization

The measure specifications are harmonized with related measures;
OR
The differences in specifications are justified

5a.1. **If this measure conceptually addresses EITHER the same measure focus OR the same target population as NQF-endorsed measure(s):**
Are the measure specifications completely harmonized?

5a.2. **If the measure specifications are not completely harmonized, identify the differences, rationale, and impact on interpretability and data collection burden.**

5b. Competing Measures

The measure is superior to competing measures (e.g., is a more valid or efficient way to measure);
OR
Multiple measures are justified.

5b.1. **If this measure conceptually addresses both the same measure focus and the same target population as NQF-endorsed measure(s):**
Describe why this measure is superior to competing measures (e.g., a more valid or efficient way to measure quality); OR provide a rationale for the additive value of endorsing an additional measure. (Provide analyses when possible.)

Appendix

A.1 Supplemental materials may be provided in an appendix. All supplemental materials (such as data collection instrument or methodology reports) should be organized in one file with a table of contents or bookmarks. If material pertains to a specific submission form number, that should be indicated. Requested information should be provided in the submission form and required attachments. There is no guarantee that supplemental materials will be reviewed.
Attachment **Attachment:** Appendix_3_23_16.pdf

Contact Information

Co.1 **Measure Steward (Intellectual Property Owner):** University Hospitals Cleveland Medical Center
Co.2 **Point of Contact:** Lawrence, Kleinman, drlarrykleinman@gmail.com, 617-669-3357-
Co.3 **Measure Developer if different from Measure Steward:** Collaboration for Pediatric Quality Measures (CAPQuaM)

Co.4 **Point of Contact:** Lawrence, Kleinman, drlarrykleinman@gmail.com, 617-669-3357-

Additional Information

Ad.1 **Workgroup/Expert Panel involved in measure development**

Provide a list of sponsoring organizations and workgroup/panel members' names and organizations. Describe the members' role in measure development.

Role: Expert Panelists
Jill Mhyre – University of Michigan
Kimberly Gregory – Women's Healthcare Quality and Performance Improvement at Cedars – Sinai
Siobhan Dolan – Einstein/Montifiore
William Grobman – Northwester
Peter Nielsen – Clinical operations, OB/GYN consultant to Surgeon General
Aaron Caughey – Oregon Health & Science University
Martha Cook Carter – FamilyCare (WomenCare, Inc)., Scott Depot, WV
Joyce Troxler – Gila Regional Medical Center
Gary Dildy – Baylor College of Medicine, Texas Children's Hospital

ROLE: Steering Committee and Investigator
Wilson Pace, MD American Academy of Family Physicians – DARTNET Institute - University of Colorado
Lynn Olson, PhD American Academy of Pediatrics
Christina Bethell, PhD, MBA, MPH Child and Adolescent Health Measurement Initiative, Johns Hopkins University (Previous OHSU)
Elizabeth Howell, MD Icahn School of Medicine at Mount Sinai
Harold Kaplan, MD Icahn School of Medicine at Mount Sinai
Lawrence Kleinman, MD, MPH Icahn School of Medicine at Mount Sinai
Rebecca Anderson Mount Sinai Medical Center
Ira Nash, MD Northshore - Long Island Jewish Medical Center (previous Mount Sinai School of Medicine)
Eyal Shemesh, MD Icahn School of Medicine at Mount Sinai
Mary Barton, MD National Committee on Quality Assurance
Charles Homer, MD, MPH US Department of HHS (previous National Institute for Child Health Quality)
Marla Clayman, PhD American Institutes for Research (previous Northwestern University)
Foster Gesten, MD New York State Dept. of Health, Office of Health Insurance Programs
Jerod M. Loeb, PhD The Joint Commission
Robert Rehm National Committee for Quality Assurance
Steve Kairys, MD American Academy of Pediatrics/QuIIN
Erin DuPree, MD The Joint Commission (previous: Icahn School of Medicine at Mount Sinai)
David Baker The Joint Commission (Previous Northwestern)
Beverley Johnson, BSN* Institute for Patient- and Family-Centered Care
Doris Peter Consumers Union

ROLE: Senior Advisory Board Member and Investigator
Shoshanna Sofaer, DrPH American Institutes for Research (previous CUNY Baruch)
Lynne Richardson, MD Icahn School of Medicine at Mount Sinai
Ian Holzman, MD Icahn School of Medicine at Mount Sinai
Marilyn Kacica, MD New York State Dept. of Health, Division of Family Health

ROLE: Senior Advisory Board Member
Marc Lashley, MD Allied Pediatrics
Gary Mirkin, MD Allied Pediatrics
Barbara Kupferman, RN, CMCN, LNC AmeriChoice by United Healthcare
L. Gregory Pawlson, MD, MPH SE Healthcare QUALITY Consultants (previous Blue Cross Blue Shield)
John Santa, MD* (previous Consumers Union)
John Clarke, MD ECRI, PA Patient Safety Authority
Scott Breidbart, MD Empire Blue Cross Blue Shield/ Anthem
Jeff Terry, MBA GE Healthcare
Robert St. Peter, MD, MPH Kansas Health Institute

Ruth Stein, MD Montefiore Children's Hospital
Arthur Aufses, MD Icahn School of Medicine at Mount Sinai
Eric Rose, MD Icahn School of Medicine at Mount Sinai
Wendy Brennan, MS National Alliance on Mental Illness of New York City (NAMI - NYC Metro)
Laurel Pickering, MPH Northeast Business Group on Health
Tony Hope, BM BCh, PhD Oxford University, The Ethox Center
Martin Hatlie, JD Partnership 4 Pt. Safety, Consumers Advancing Patient Safety
Paul Wise, MD Stanford University
Mark Chassin, MD, MPH, MPP The Joint Commission
Ira Nash, MD Icahn School of Medicine at Mount Sinai
ROLE: Investigators/Key Staff
Rusty McLouth, MS AAFP
Kasey McCracken, MPH CAHMI
Elise Barrow, MPH Icahn School of Medicine at Mount Sinai
Kasey Coyne, BA Icahn School of Medicine at Mount Sinai
Natalia Egorova, PhD Icahn School of Medicine at Mount Sinai
Elizabeth Howell, MD Icahn School of Medicine at Mount Sinai
Harold Kaplan, MD Icahn School of Medicine at Mount Sinai
Melissa Manice, MPH Icahn School of Medicine at Mount Sinai
Joshua Oppenheimer, BA Icahn School of Medicine at Mount Sinai
Eve Oster, MD Pediatric Partners of Gwinnett (previous Icahn School of Medicine at Mount Sinai)
Daniel O'Sullivan, BS University of Virginia (previous Icahn School of Medicine at Mount Sinai)
Michael Parides, PhD Icahn School of Medicine at Mount Sinai
Barbara Rabin, MHA Icahn School of Medicine at Mount Sinai
Carolyn Rosen, MD Icahn School of Medicine at Mount Sinai
Melissa Saperstein, MSW Icahn School of Medicine at Mount Sinai
Eyal Shemesh, MD Icahn School of Medicine at Mount Sinai
Andrew Ting, MD Icahn School of Medicine at Mount Sinai
Adam Vella, MD Icahn School of Medicine at Mount Sinai
Virginia Walther, MSW Icahn School of Medicine at Mount Sinai
Jennifer Zeitlin, SD Icahn School of Medicine at Mount Sinai, French Institute of Health and Medical Research (INSERM)
Marianne McPherson, PhD, MS NICHQ
Joseph Anarella, MPH NYS DOH, OHIP
Lee Sanders, MD, MPH Stanford University
Kasey Coyne Northwestern
Victoria Wagner New York State Health Department Office Quality Patient Safety
Laura Soloway New York State Health Department Office Quality Patient Safety
Kevin Malloy New York State Health Department Office Quality Patient Safety
Judy Stribling Icahn School of Medicine at Mount Sinai
Linda Radecki AAP
Keri Thiessen AAP – QuIIN
Laura Morris New York State Department of Health
Olivia Lindly OHSU
Kasey McCracken OHSU
Sandeep Sharma Icahn School of Medicine at Mount Sinai
Ann Nevar UH Rainbow Babies and Children's Hospital/ CAPQuaM
Suzanne Lo UH Rainbow Babies and Children's Hospital/ CAPQuaM

Amy Balbierz Icahn School of Medicine at Mount Sinai
Samantha Raymond Icahn School of Medicine at Mount Sinai
Allisyn Vachon Icahn School of Medicine at Mount Sinai
Shannon Weber Icahn School of Medicine at Mount Sinai
ROLE: Contractors
Mary Fialkowski, MBA The Joint Commission
Linda Hanold, MHSA The Joint Commission
Elvira Ryan, MBA, BSN, RN The Joint Commission

Sharon Sprenger, RHIA, CPHQ, MPA The Joint Commission Ann Watt, MBA, RHIA The Joint Commission Amy Slott The Joint Commission Denise Krusenoski The Joint Commission Elvira Ryan The Joint Commission Tasha Mearday The Joint Commission
Measure Developer/Steward Updates and Ongoing Maintenance Ad.2 **Year the measure was first released:** Ad.3 **Month and Year of most recent revision:** Ad.4 **What is your frequency for review/update of this measure?** Ad.5 **When is the next scheduled review/update for this measure?**
Ad.6 **Copyright statement:** Ad.7 **Disclaimers:**
Ad.8 **Additional Information/Comments:**

MEASURE WORKSHEET

This document summarizes the evaluation of the measure as it progresses through NQF's Consensus Development Process (CDP). The information submitted by measure developers/stewards is included after the Brief Measure Information, Preliminary Analysis, and Pre-meeting Public and Member Comments sections.

To navigate the links in the worksheet: Ctrl + click link to go to the link; ALT + LEFT ARROW to return

Brief Measure Information

NQF #: 2902
Measure Title: Contraceptive Care - Postpartum
Measure Steward: US Office of Population Affairs
Brief Description of Measure: Among women ages 15 through 44 who had a live birth, the percentage that is provided:
1) A most effective (i.e., sterilization, implants, intrauterine devices or systems (IUD/IUS)) or moderately (i.e., injectables, oral pills, patch, ring, or diaphragm) effective method of contraception within 3 and 60 days of delivery.
2) A long-acting reversible method of contraception (LARC) within 3 and 60 days of delivery.
Two time periods are proposed (i.e., within 3 and within 60 days of delivery) because each reflects important clinical recommendations from the U.S. Centers for Disease Control and Prevention (CDC) and the American College of Obstetricians and Gynecologists (ACOG). The 60-day period reflects ACOG recommendations that women should receive contraceptive care at the 6-week postpartum visit. The 3-day period reflects CDC and ACOG recommendations that the immediate postpartum period (i.e., at delivery, while the woman is in the hospital) is a safe time to provide contraception, which may offer greater convenience to the client and avoid missed opportunities to provide contraceptive care.
Developer Rationale: Contraception is a highly effective clinical preventive service intended to prevent teen and unintended pregnancy and space births. The type of contraceptive method used by a woman is strongly associated with her risk of unintended pregnancy [1, 2]. The most effective methods (sterilization and the long-acting reversible methods of IUDs and implants) have a failure rate that is less than 1% per year under typical use; the moderately effective methods (shot, pill, patch, ring and diaphragm) have a typical failure rate of 6-12% per year; the least effective methods have a typical failure rate of 18-28%; and 85% of women will become pregnant in a year if they do not use any contraceptive method at all [2]. The proposed performance measure is based on the fact that some contraceptive methods are more effective than others in preventing unintended pregnancy, and hence, in spacing births among postpartum women.

The policy context is increasingly supportive of efforts to provide contraception in the postpartum period. ACOG, AAP and CDC recommend the provision of contraception in the postpartum period and describe how it can be done so safely (see evidence report for details). Financial barriers to immediate postpartum LARC insertion have been removed by numerous state Medicaid programs (i.e., by reimbursing separately for LARC in the immediate postpartum period, outside of the bundled delivery payment) [3]. The cost effectiveness of this practice has been documented [4, 5], and more states and private health plans are likely to adopt this policy in the near future. Given this context, we expect that use of the proposed clinical performance measures will encourage more providers to follow ACOG (2015) and CDC (2010, 2014) recommendations to:

- During prenatal visits, counsel all pregnant women about postpartum contraception, including the possibility of obtaining LARC and other methods of contraception at delivery, i.e., in the immediate postpartum period, and the effectiveness of the different methods as well as other factors that may help a woman decide the method that is best for her [6]. They will also take steps to make LARC available in the immediate postpartum inpatient setting [7].
- At the postpartum visit, counsel all women about contraception, which includes information about the effectiveness of the different methods as well as other factors that will help a woman choose the method that is best for her [6]. They will also take steps to make LARC available on a same-day basis in the outpatient setting [7].

References

1. Mansour, D., P. Inki, and K. Gemzell-Danielsson, Efficacy of contraceptive methods: a review of the literature. Eur J Contracept Reprod Health Care, 2010. 15 Suppl 2: p. S19-31.

2. Trussell, J., Contraceptive failure in the United States. Contraception, 2011. 83(5): p. 397-404.
3. Moniz, M.H., et al., Characterization of Medicaid policy for immediate postpartum contraception. Contraception, 2015. 92(6): p. 523-31.
4. Han, L., et al., Cost-effectiveness of immediate postpartum esonogestrel implant insertion for adolescent mothers Contraception. 86: p. 290–325.
5. Rodriguez, M.I., et al., Cost-benefit analysis of state- and hospital-funded postpartum intrauterine contraception at a university hospital for recent immigrants to the United States. Contraception, 2010. 81(4): p. 304-8.
6. CDC, Providing Quality Family Planning Services: Recommendations of the CDC and the U.S. Office of Population Affairs. MMWR Recommendations and Reports, 2014. 63(4): p. 1-54.

Numerator Statement: Primary measure: Women ages 15 through 44 who had a live birth and were provided a most (sterilization, intrauterine device, implant) or moderately (pill, patch, ring, injectable, diaphragm) effective method of contraception within 3 and 60 days of delivery.
Sub-measure: Women ages 15 through 44 who had a live birth and were provided a long-acting reversible method of contraception (LARC) within 3 and 60 days of delivery.
Denominator Statement: Women ages 15 through 44 who had a live birth in a 12-month measurement year.
Denominator Exclusions: The following categories are excluded from the denominator: (1) deliveries that did not end in a live birth (i.e., miscarriage, ectopic, stillbirth or induced abortion); and (2) deliveries that occurred during the last two months of the measurement year.

Measure Type: Intermediate Clinical Outcome
Data Source: Administrative claims
Level of Analysis: Health Plan, Population : Regional

IF Endorsement Maintenance – Original Endorsement Date: Most Recent Endorsement Date:

New Measure -- Preliminary Analysis

Criteria 1: Importance to Measure and Report

1a. Evidence

1a. Evidence. The evidence requirements for a *process or intermediate outcome* measure is that it is based on a systematic review (SR) and grading of the body of empirical evidence where the specific focus of the evidence matches what is being measured.

The evidence for this intermediate outcome measure should demonstrate the evidence-based relationship between providing contraception and health outcomes.

The developer provides the following evidence for this measure:

- **Systematic Review of the evidence specific to this measure?** ☒ **Yes** ☐ **No**
- **Quality, Quantity and Consistency of evidence provided?** ☒ **Yes** ☐ **No**
- **Evidence graded?** ☐ **Yes** ☒ **No**

Summary of evidence:

- The developer provides a diagram of the relationship between the intermediate outcome (use of long-acting, reversible contraception) and health outcomes (reduction in unintended pregnancy
- The developer describes three ungraded guidelines from CDC/OPA, ACOG and AAP recommending counseling about contraception and focusing on most effective methods.
- Three systematic reviews (SR) on the efficacy of contraceptive methods are presented and include one SR specific to post-partum contraception.
- QQC is provided for the SR, though the quality is implied by the large number of RCTs included in the SR.
- Conclusions of SRs:
  - (2011) “the most effective methods (LARC and sterilization) have a failure rate that is less than 1% per

year under typical use; the moderately effective methods (shot, PPR, diaphragm) have a typical failure rate of 6-12% per year"

- (2010) "the review broadly confirmed the hierarchy of contraceptive effectiveness in descending order as: (1) female sterilization, long-acting hormonal contraceptives (LNG-IUS and implants); (2) Cu-IUDs with_300 mm2 surface area; (3) Cu-IUDs with5300 mm2 surface area and short-acting hormonal contraceptives (injectables, oral contraceptives, the patch and vaginal ring), and (4) barrier methods and natural methods."
- (2015) A meta-analysis showed that IUC use at six months was more likely with immediate insertion than with standard insertion (OR 2.04; 95% CI 1.10 to 4.09) although the expulsion rate was higher.

Guidance from the Evidence Algorithm

Process measure (Box 1) → Systematic review (Box 3) → QQC present (Box 4) →SR concludes high quality evidence

Question for the Committee:

- *What is the relationship of this measure to patient outcomes?*
- *How strong is the evidence for this relationship?*
- *Is the evidence directly applicable to the process of care being measured?*

Preliminary rating for evidence: ☒ **High** ☐ **Moderate** ☐ **Low** ☐ **Insufficient**

1b. Gap in Care/Opportunity for Improvement and 1b. disparities

1b. Performance Gap. The performance gap requirements include demonstrating quality problems and opportunity for improvement.

The developers describe the data found during measure testing:

	Iowa Medicaid	Louisiana Medicaid
By 3 days	10.6%	9.7%
By 60 days	41.6-42.6%	42.5 – 47.3%

Disparities

- Developers report that "An analysis of data from the Pregnancy Risk Assessment Monitoring System (PRAMS), 2011-2012, suggests that there are statistically significant differences by age, marital status and some income categories for use of most and moderately effective methods and LARC methods. However, there were no significant differences by race/ethnicity, and most income categories."

Questions for the Committee:

- *Is there opportunity for improvement in performance?*
- *Is there a gap in care that warrants a national performance measure?*
- *If no disparities information is provided, are you aware of evidence that disparities exist in this area of healthcare?*

Preliminary rating for opportunity for improvement: ☒ **High** ☐ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 1: Importance to Measure and Report (including 1a, 1b, 1c)

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability

2a1. Reliability Specifications

2a1. Specifications requires the measure, as specified, to produce consistent (reliable) and credible (valid) results about the quality of care when implemented.

Data source(s): Administrative claims

Specifications:

- The numerator is the number of women who were provided a most (sterilization, intrauterine device, implant) or moderately (pill, patch, ring, injectable, diaphragm) effective method of contraception within 3 and 60 days of delivery.
- The denominator is women ages 15 through 44 who had a live birth in a 12-month measurement year. In a Medicaid population, this includes women who were enrolled from the date of delivery to 60 days postpartum.
- The developers tested the measure specifications based on 2014 codes, but have also included the codes needed to calculate the measure using ICD-10 and 2015 NDC codes. Both sets of codes are attached. The goal was to convert the measure to a new code set, fully consistent with the intent of the original measure.
- Exclusions include: (1) deliveries that did not end in a live birth (i.e., miscarriage, ectopic, stillbirth or induced abortion); and (2) deliveries that occurred during the last two months of the measurement year.
- The developers suggest stratification by age, so that adolescents can be examined separately from adult women—this stratification is for purposes of QI but not as a method of risk adjustment.
- A calculation algorithm is included.

Questions for the Committee:

- *Are all the data elements clearly defined? Are all appropriate codes included?*
- *Is the logic or calculation algorithm clear?*
- *Is it likely this measure can be consistently implemented?*

2a2. Reliability Testing Testing attachment

2a2. Reliability testing demonstrates if the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise enough to distinguish differences in performance across providers.

SUMMARY OF TESTING

Reliability testing level ☒ **Measure score** ☐ **Data element** ☐ **Both**

Reliability testing performed with the data source and level of analysis indicated for this measure ☒ **Yes** ☐ **No**

Method(s) of reliability testing

- Testing was done using Medicaid datasets from Iowa and Louisiana. Testing was performed at the health plan level and public health region levels of analysis.
- The developers describe their signal-to-noise testing approach using the Intraclass Correlation (ICC) with a Spearman-Brown prophesy formula to estimate reliability. This is an appropriate method for estimating the reliability of the measure score. Generally, reliability estimates > 0.70 are considered acceptable.

Results of reliability testing

The developers provide testing results in a table for ICC and reliability.

From the testing, the developers conclude:

- For the 60 day most/moderately effective measure, we would recommend that regions or plans have at least 862 patient cases for reporting rates to maintain >.70 reliability, and 3,324 cases to maintain >.90 reliability.

- For the 60-day LARC measure, we would recommend that regions or plans have at least 553 patient cases for reporting rates to maintain >.70 reliability, and 2,134 cases to maintain >.90 reliability.

- For the 3-day most/moderately effective and LARC measures, we looked at the set of ICC values for regions and plans among the 15-44 age group only (we do not recommend stratifying by teen/adult at this time). We would recommend that regions or plans have at least 1,058 patient cases for reporting rates to maintain >.70 reliability, and 4,082 cases to maintain >.90 reliability.

Guidance from the Reliability Algorithm

Precise specifications (Box 1)→ empirical reliability testing (Box 2) → testing of computed measure score (Box 4)→ appropriate method (Box 5) → high certainty that scores or reliable within the reporting rates recommended by the developer (Box 6a) → high

Questions for the Committee:

- *Is the test sample adequate to generalize for widespread implementation?*
- *Do the results demonstrate sufficient reliability so that differences in performance can be identified?*

Preliminary rating for reliability: ☒ **High** ☐ **Moderate** ☐ **Low** ☐ **Insufficient**

2b. Validity

2b1. Validity: Specifications

2b1. Validity Specifications. This section should determine if the measure specifications are consistent with the evidence.

Specifications consistent with evidence in 1a. ☒ **Yes** ☐ **Somewhat** ☐ **No**

Question for the Committee:

- *Are the specifications consistent with the evidence?*

2b2. Validity testing

2b2. Validity Testing should demonstrate the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality.

SUMMARY OF TESTING

Validity testing level ☒ **Measure score** ☐ **Data element testing against a gold standard** ☐ **Both**

Method of validity testing of the measure score:

☒ **Face validity only**

☐ **Empirical validity testing of the measure score**

Validity testing method:

- Systematic assessment of face validity from an eight member panel of experts:

Validity testing results:

The panel rated the validity of the measure on a 1-5 scale:

- The measure assessing provision of most/moderately effective methods will provide an accurate reflection of quality and can be used to distinguish good and poor quality in contraceptive services (mean rating = 4.22).
- Although the majority of panel members agreed or strongly agreed that the LARC post-partum measure will provide an accurate reflection of quality,
- Some concerns were raised about the validity in breastfeeding women (mean rating = 3.78, with one member strongly disagreeing due to the breastfeeding issue).

Staff note: Because only face validity was assessed rather than empirical validity testing, the highest rating possible for validity is MODERATE.

Questions for the Committee:

- *Is the test sample adequate to generalize for widespread implementation?*
- *Do the results demonstrate sufficient validity so that conclusions about quality can be made?*
- *Do you agree that the score from this measure as specified is an indicator of quality?*

2b3-2b7. Threats to Validity

2b3. Exclusions:

The developers describe the frequency of two exclusions in the testing data sets:

	IOWA	***LOUISIANA***
Deliveries that did not end in a live birth (miscarriage, ectopic, stillbirth or induced abortion)	240 (1.6%)	1481 (3.5%)
Live births that occurred in the last 2 months of the measurement year	2167 (14.6%)	7142 (16.8%)

Questions for the Committee:

- *Are the exclusions consistent with the evidence?*
- *Are any patients or patient groups inappropriately excluded from the measure?*
- *Are the exclusions/exceptions of sufficient frequency and variation across providers to be needed (and outweigh the data collection burden)?*

2b4. Risk adjustment: **Risk-adjustment method** ☒ **None** ☐ **Statistical model** ☐ **Stratification**

2b5. Meaningful difference *(can statistically significant and clinically/practically meaningful differences in performance measure scores can be identified)*:

- The developers provide a table with mean results and ranges of values for the two datasets and by age.
- The developers conclude "There are large differences in rates across both levels for the 60-day measures of postpartum contraception. For example, the rates for most and moderately effective methods ranged by region in Iowa (from 35% to 48%) and Louisiana (from 35% to 49%), and across health plans in Louisiana (from 36% to 47%). There were also substantial differences in provision of LARC methods across regions in Iowa (from 8% to 17%) and Louisiana (from4% to 13%), and across health plans in Louisiana (from 4% to 16%). These differences suggest that it will be possible to identify meaningful differences in performance across measured entities for the 60-day measures. "

Question for the Committee:

- *Does this measure identify meaningful differences about quality?*

2b6. Comparability of data sources/methods: NA

2b7. Missing Data

- The developers note that "The data source for this measure is claims data. Due to the nature of claims data (i.e., for billing purposes), there is typically very little missing data; further, it is difficult to ascertain when claims data is or is not missing."

Guidance from the algorithm:
Consistent with specifications (Box 1) →potential threats to validity addressed (Box 2) →face validity systematically assessed (Box 4) → substantial agreement (Box 5) → moderate (in the absence of empirical validity testing, moderate is the highest rating possible)

Preliminary rating for validity: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments
Criteria 2: Scientific Acceptability of Measure Properties (including all 2a, 2b, and 2d)

Criterion 3. Feasibility

3. Feasibility is the extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

- This measure is based on administrative claims. Claims data are generally thought to be of minimal burden and highly feasible.

Questions for the Committee:

- *Are the required data elements routinely generated and used during care delivery?*
- *Are the required data elements available in electronic form, e.g., EHR or other electronic sources?*
- *Is the data collection strategy ready to be put into operational use?*

Preliminary rating for feasibility: ☒ **High** ☐ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments
Criteria 3: Feasibility

Criterion 4: Usability and Use

4. Usability and Use evaluate the extent to which audiences (e.g., consumers, purchasers, providers, policymakers) use or could use performance results for both accountability and performance improvement activities.

Current uses of the measure
Publicly reported? ☐ **Yes** ☒ **No**

Current use in an accountability program? ☐ **Yes** ☐ **No**
OR
Planned use in an accountability program? ☒ **Yes** ☐ **No**

Accountability program details

- Public reporting planned
- Centers for Medicaid and CHIP Services (CMCS) has developed the Maternal and Infant Health initiative to improve the quality of maternity care and birth outcomes, and to measure how care is delivered to women.

CMCS is funding 13 states to collect and report data on this new contraception measure on a developmental basis (pending NQF endorsement), to help states track the provision of postpartum contraception and to drive changes in care practices and delivery.

Improvement results NA

Unexpected findings (positive or negative) during implementation none identified

Potential harms

- The developers point out "The one issue that has arisen as a potential concern is that the measure may lead to coercive practices in which women are not offered a free choice of methods and are pressured into using the most and moderately effective, or LARC, methods (Gold 2014, Dehlendorf 2015)." The developers provide an explanation of why they do not believe this will be a significant issue in using this measure.

Feedback: The Measures Application Partnership (MAP) recommended use of the contraceptive measures in CMCS' core adult and child quality measures, if they are endorsed by NQF.

***Questions for the Committee**:*

- *How can the performance results be used to further the goal of high-quality, efficient healthcare?*
- *Do the benefits of the measure outweigh any potential unintended consequences?*

Preliminary rating for usability and use: ☒ **High** ☐ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 4: Usability and Use

Criterion 5: Related and Competing Measures

Related or competing measures

- 1517: Prenatal & Postpartum Care (PPC)
- 2903 Contraceptive Care – Most and Moderately Effective Methods
- 2904 Contraceptive Care – Access to LARC

Harmonization

- This measure (#2902) considers contraceptive care for the same population addressed in the NCQA measure on prenatal and postpartum care (PPC) (NQF#1517), although the measures address different types of services. The developers state that they have aligned the contraceptive measure with the PCC measure to the extent possible, with regard to identifying the population of women with live births.
- Measures 2903 and 2904 are from the same developer and harmonized.

Pre-meeting public and member comments

- Planned Parenthood Federation of America, the nation's leading provider of women's reproductive healthcare, supports the endorsement of the proposed measures. Contraception is an important and effective preventive service to reduce unintended pregnancy as well as improve birth spacing and family planning. PPFA provided de-identified data included in the application to demonstrate the reliability and validity of the measures as well the

feasibility of using them for quality improvement. Currently, PPFA has already begun using a developmental version of these measures for quality improvement and looks forward to incorporate NQF endorsed measure into its portfolio of internal quality improvement work. National endorsement of these new performance measures on contraceptive care aligns with the April 2015 call by the Institute of Medicine for standardized metrics that include measuring contraceptive use to support reducing unintended pregnancy. Further, these will be the first nationally endorsed measures on contraceptive care, providing important tools to all providers who serve women of reproductive age.

NATIONAL QUALITY FORUM—Evidence (subcriterion 1a)

Measure Number (*if previously endorsed*)**:** Click here to enter NQF number
Measure Title: Contraceptive Care – Postpartum
IF the measure is a component in a composite performance measure, provide the title of the Composite Measure here: Click here to enter composite measure #/ title

Date of Submission: 2/15/2016

Instructions

- *For composite performance measures:*
  - *A separate evidence form is required for each component measure unless several components were studied together.*
  - *If a component measure is submitted as an individual performance measure, attach the evidence form to the individual measure submission.*
- Respond to all questions as instructed with answers immediately following the question. All information needed to demonstrate meeting the evidence subcriterion (1a) must be in this form. An appendix of *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 10 pages (*incudes questions/instructions*; minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.

Note: The information provided in this form is intended to aid the Steering Committee and other stakeholders in understanding to what degree the evidence for this measure meets NQF's evaluation criteria.

1a. Evidence to Support the Measure Focus
The measure focus is evidence-based, demonstrated as follows:

- Health outcome: [3] a rationale supports the relationship of the health outcome to processes or structures of care. Applies to patient-reported outcomes (PRO), including health-related quality of life/functional status, symptom/symptom burden, experience with care, health-related behavior.
- Intermediate clinical outcome: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured intermediate clinical outcome leads to a desired health outcome.
- Process: [5] a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured process leads to a desired health outcome.
- Structure: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured structure leads to a desired health outcome.
- Efficiency: [6] evidence not required for the resource use component.

Notes

3. Generally, rare event outcomes do not provide adequate information for improvement or discrimination; however, serious reportable events that are compared to zero are appropriate outcomes for public reporting and quality improvement.
4. The preferred systems for grading the evidence are the U.S. Preventive Services Task Force (USPSTF) grading definitions and methods, or Grading of Recommendations, Assessment, Development and Evaluation (GRADE) guidelines.
5. Clinical care processes typically include multiple steps: assess → identify problem/potential problem → choose/plan intervention (with patient input) → provide intervention → evaluate impact on health status. If the measure focus is one step in such a multistep process, the step with the strongest evidence for the link to the desired outcome should be selected as the focus of measurement. Note: A measure focused only on collecting PROM data is not a PRO-PM.
6. Measures of efficiency combine the concepts of resource use and quality (see NQF's Measurement Framework: Evaluating Efficiency Across Episodes of Care; AQA Principles of Efficiency Measures).

1a.1.This is a measure of: (*should be consistent with type of measure entered in De.1*)
Outcome

☐ Health outcome: Click here to name the health outcome

☐Patient-reported outcome (PRO): Click here to name the PRO

PROs include HRQoL/functional status, symptom/symptom burden, experience with care, health-related behaviors

☒ Intermediate clinical outcome (*e.g., lab value*): Contraceptive use

☐ Process: Click here to name the process

☐ Structure: Click here to name the structure

☐ Other: Click here to name what is being measured

_____

HEALTH OUTCOME/PRO PERFORMANCE MEASURE *If not a health outcome or PRO, skip to 1a.3*

1a.2. Briefly state or diagram the path between the health outcome (or PRO) and the healthcare structures, processes, interventions, or services that influence it.

Not a health outcome or PRO.

1a.2.1. State the rationale supporting the relationship between the health outcome (or PRO) to at least one healthcare structure, process, intervention, or service (*i.e., influence on outcome/PRO*).

Note: For health outcome/PRO performance measures, no further information is required; however, you may provide evidence for any of the structures, processes, interventions, or service identified above.

_____

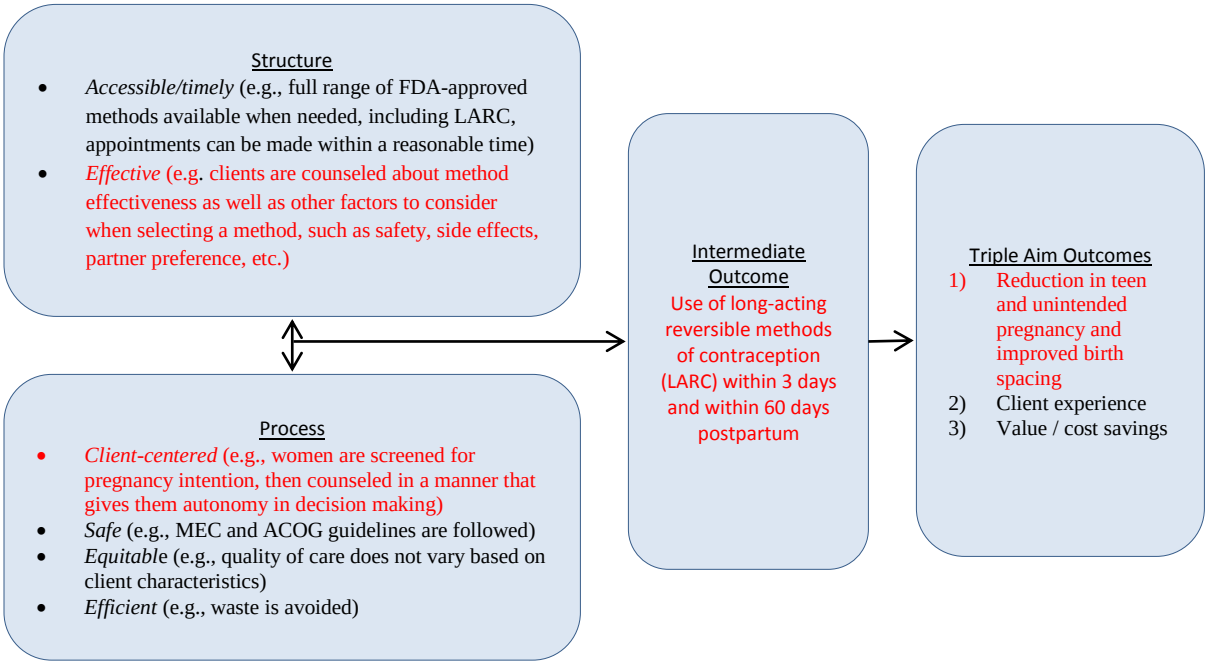
INTERMEDIATE OUTCOME, PROCESS, OR STRUCTURE PERFORMANCE MEASURE

1a.3. Briefly state or diagram the path between structure, process, intermediate outcome, and health outcomes. Include all the steps between the measure focus and the health outcome.

The diagram below illustrates the steps between the structure and process that influence the intermediate health outcome, and how the intermediate health outcome in turns influences the longer-term outcomes. The text highlighted in red shows the primary relationships that will be affected by use of the proposed measure: (a) increased use of the most and moderately effective methods of contraception will influence rates of unintended pregnancy; and (b) appropriate counseling of a client can lead to increased use of the most and moderately effective methods of contraception.

The type of contraceptive method used by a woman is strongly associated with her risk of unintended pregnancy. The most effective methods (sterilization and the long-acting reversible contraceptive [LARC] methods of intrauterine devices and implants) have a failure rate that is less than 1% per year under typical use; the moderately effective methods (shot, oral pills, patch, ring, and diaphragm) have a typical failure rate of 6-12% per year; the least effective methods have a typical failure rate of 18-28%; and if no method is used then 85 of every 100 women will become pregnant in a year (Trussell 2011).

The measure is secondarily supported by evidence that the way in which contraceptive counseling is offered (e.g., increased screening of clients for reproductive intention; the provision of client-centered counseling, which includes providing information about and ready access to the most and moderately effective methods of contraception; and ready access to all methods of contraception, ideally on a same-day basis) will lead to increased use of the most and moderately effective methods of contraception (i.e., the intermediate outcome).



1a.3.1. What is the source of the systematic review of the body of evidence that supports the performance measure?

☒ Clinical Practice Guideline recommendation – ***complete sections*** 1a.4***, and*** 1a.7

☐ US Preventive Services Task Force Recommendation – ***complete sections*** 1a.5 ***and*** 1a.7

☒ Other systematic review and grading of the body of evidence (*e.g., Cochrane Collaboration, AHRQ Evidence Practice Center*) – ***complete sections*** 1a.6 ***and*** 1a.7

☐ Other – ***complete section*** 1a.8

Please complete the sections indicated above for the source of evidence. You may skip the sections that do not apply.

1a.4. CLINICAL PRACTICE GUIDELINE RECOMMENDATION

1a.4.1. Guideline citation (*including date*) and **URL for guideline** (*if available online*):

Clinical recommendations (from both government sources and professional organizations) are the best source of evidence about the relationship between contraceptive counseling and increased use of the most and moderately effective methods of contraception (see diagram above).

CDC/OPA (2014). Providing Quality Family Planning Services (QFP): Recommendations of CDC and the US Office of Population Affairs, MMWR Recommendations and Reports, April 24, 2014. http://www.cdc.gov/mmwr/preview/mmwrhtml/rr6304a1.htm

American College of Obstetricians and Gynecologists (ACOG), Committee on Gynecologic Practice. Increasing access to contraceptive implants and intrauterine devices to reduce unintended pregnancy. Committee Opinion Number 642; October 2015.

ACOG *Long-acting reversible contraception: Implants and intrauterine devices*, in *Practice Bulletin*. 2015 (reaffirmed), American College of Obstetricians and Gynecologists: Washington, DC. p. 1-13.

The American Academy of Pediatrics (AAP) (2014). Contraception for Adolescents. Pediatrics, 134:e1244–e1256.

1a.4.2. Identify guideline recommendation number and/or page number and **quote verbatim, the specific guideline recommendation**.

"Providers are encouraged to present information on potential reversible methods of contraception by using a tiered approach (i.e., presenting information on the most effective methods first, before presenting information on less effective methods). This information should include an explanation that long-acting reversible contraceptive methods are safe and effective for most women, including those who have never given birth and adolescents. Information should be tailored and presented to ensure a client-centered approach. It is not appropriate to omit presenting information on a method solely because the method is not available at the service site. If not all methods are available at the service site, it is important to have strong referral links in place to other providers to maximize opportunities for clients to obtain their preferred method that is medically appropriate."
Source: CDC/OPA (2014). *Providing Quality Family Planning Services*, page 8 and Appendix B

"For all women at risk of unintended pregnancy, obstetrician-gynecologists should provide counseling on all contraceptive options, including implants and IUDs. Long-acting reversible contraception methods require a single action of motivation for long-term use, eliminating adherence and user dependence from the effectiveness equation. These top-tier methods share the highest continuation rates of all contraceptives, which is one of the most important factors in contraceptive success."
Source: ACOG (2015), page 1.

"The immediate postpartum period is a particularly favorable time for IUD or implant insertion. Women who have recently given birth are often highly motivated to use contraception, they are known not to be pregnant and the hospital setting offers convenience for both the patient and the health care provider." ACOG (2015 Practice Bulletin), page 4.

"Contraceptive methods most commonly used by adolescents are listed below, ordered from most to least effective, starting with long-acting reversible contraception (LARC); implants and IUDs. *Pediatricians are encouraged to counsel adolescents in that order, discussing the most effective contraceptive methods first*." ACOG (2014), page e1246.

1a.4.3. Grade assigned to the quoted recommendation <u>with definition</u> of the grade:
Not applicable

1a.4.4. Provide all other grades and associated definitions for recommendations in the grading system. (*Note: If separate grades for the strength of the evidence, report them in section 1a.7.*)
Not applicable

1a.4.5. Citation and URL for methodology for grading recommendations (*if different from 1a.4.1*)**:**
Not applicable

1a.4.6. If guideline is evidence-based (rather than expert opinion), are the details of the quantity, quality, and consistency of the body of evidence available (e.g., evidence tables)?

☒ Yes → ***complete section*** 1a.7
☐ No → report on another systematic review of the evidence in sections 1a.6 and 1a.7; if another review does not exist, ***provide what is known from the guideline review of evidence in*** 1a.7

1a.5. UNITED STATES PREVENTIVE SERVICES TASK FORCE RECOMMENDATION
1a.5.1. Recommendation citation (*including date*) and **URL for recommendation** (*if available online*):

Not applicable

1a.5.2. Identify recommendation number and/or page number and **quote verbatim, the specific recommendation**.

Not applicable

1a.5.3. Grade assigned to the quoted recommendation with definition of the grade:

Not applicable

1a.5.4. Provide all other grades and associated definitions for recommendations in the grading system. (*Note: the grading system for the evidence should be reported in section 1a.7.*)

Not applicable

1a.5.5. Citation and URL for methodology for grading recommendations (*if different from 1a.5.1*)**:**

Not applicable

Complete section 1a.7

_____

1a.6. OTHER SYSTEMATIC REVIEW OF THE BODY OF EVIDENCE

1a.6.1. Citation (*including date*) and **URL** (*if available online*):

Two systematic literature reviews are the best source of evidence about the relationship between use of long-acting reversible methods of contraception (LARC) and unintended pregnancy (see diagram in 1a.3, above). A third systematic review focused on the provision of LARC methods in the immediate postpartum period.

1. The first review was led by Professor James Trussell from Princeton University, which is repeated on an ongoing basis and published in a handbook entitled "Contraceptive Technology". The Trussell analyses serve as the primary source of information about contraceptive failure rates, and are cited by the World Health Organization, CDC, and leading professional associations in the U.S. and in other countries. Trussell used two sources of data when estimating contraceptive failure. The first was published research, which comprised results from clinical trials and surveys. The second source was the CDC's National Survey of Family Growth (NSFG), which was used to estimate *typical* use rates using data from a nationally representative sample of users.

- Trussell J (2011). Contraceptive efficacy. In: Hatcher RA, Trussell J, Nelson AL, Cates W, Kowal D, editors. Contraceptive technology: twentieth revised edition. New York: Ardent Media; 2011, pp. 777–861. This was subsequently summarized in: Trussell J (2011). Contraceptive failure in the United States. Contraception; 83(5):397-404.

- WHO/Department of Reproductive Health and Research & Johns Hopkins Bloomberg School of Public Health/Center for Communication Programs (2011). Family Planning: A Global Handbook for Providers. Baltimore and Geneva: CCP and WHO.

2. The second review was conducted by Mansour et al in 2010. They search Medline and Embase from January 1990 to February 2008 for publications reporting contraceptive failure rates.

    - Mansour D, Inki P, Gemzell-Danielsson K (2010). Efficacy of contraceptive methods: A review of the literature. The European Journal of Contraception and Reproductive Health Care, 15:4-16.

3. A recent Cochrane systematic review examined the outcomes of IUD insertion immediately after placement delivery (within 10 minutes). Randomized clinical trials published through April 1, 2015 were identified in the following databases: PubMed, CENTRAL, POPLINE, Web of Science, EMBASE, LILACS, ClinicalTrials.gov, and ICTRP.

    - Lopez, L.M., et al., *Immediate postpartum insertion of intrauterine device for contraception.* Cochrane Database Syst Rev, 2015. **6**: p. CD003036.

1a.6.2. Citation and URL for methodology for evidence review and grading (*if different from 1a.6.1*)**:**

Evidence of contraceptive effectiveness & its impact on unintended pregnancy

- See 1a.6.1 above

Evidence of effectiveness of counseling or other interventions to affect patients' choice of method

- Zapata LB, Tregear SJ, Curtis KM, Tiller M, Pazol K, Mautone-Smith N, Gavin LE (2015). Impact of Contraceptive Counseling in Clinical Settings: A Systematic Review. Am J Prev Med. 2015 Aug;49(2 Suppl 1):S31-45.

Complete section 1a.7

_____

1a.7. FINDINGS FROM SYSTEMATIC REVIEW OF BODY OF THE EVIDENCE SUPPORTING THE MEASURE

If more than one systematic review of the evidence is identified above, you may choose to summarize the one (or more) for which the best information is available to provide a summary of the quantity, quality, and consistency of the body of evidence. Be sure to identify which review is the basis of the responses in this section and if more than one, provide a separate response for each review.

1a.7.1. What was the specific structure, treatment, intervention, service, or intermediate outcome addressed in the evidence review?

Evidence of contraceptive effectiveness & its impact on unintended pregnancy

The studies examining contraceptive efficacy and effectiveness considered the impact of use of specific contraceptive methods on risk of pregnancy (i.e., contraceptive failure). Pregnancy risk can be assessed either through life table analyses (usually through 12 months) that show the percentage of women who become pregnant, or the score on the Pearl Index. The Pearl Index is a commonly used technique for reporting the effectiveness of a birth control method in clinical trials, and estimates the number of unintended pregnancies over a period of exposure (e.g. 100 women over one year of use, or 10 women over 10 years). Contraceptive failure rates are reported for *perfect use* and *typical use*. Perfect use reflects how effective methods can be in preventing pregnancy when used consistently and correctly according to instructions. Typical use reflects how effective methods are for the average person who does not always use methods correctly or consistently. Pregnancy rates during typical use of adherence-dependent methods (such as the oral pill) generally vary widely for different groups using the same method, primarily due to differences in the propensity to use the method perfectly. The review by Lopez et al (2015) focused on immediate postpartum insertion of IUDs (within 10 minutes) compared immediate insertion to insertion at other postpartum times. Key outcomes were expulsion and method use.

Evidence of effectiveness of counseling or other interventions to affect patients' choice of method

The systematic review underpinning the CDC-OPA recommendation on contraceptive counseling used an analytic framework that considered the impact of providing contraceptive counseling and/or education on short (e.g., client knowledge, attitudes), medium (e.g., selection of more effective methods, correct and consistent use) and long-term (unintended pregnancy) outcomes (Zapata 2015).

1a.7.2. Grade assigned for the quality of the quoted evidence with definition of the grade:

Evidence of contraceptive effectiveness & its impact on unintended pregnancy

While the quality of the studies was not graded in either the Trussell (2011) or Mansour (2010) review, they were primarily comprised of randomized controlled trials. The Lopez (2015) review applied principles from GRADE (Grades of Recommendation, Assessment, Development and Evaluation) to assess the quality of evidence as shown below, and found the body of evidence to be of moderate quality:

- **High quality:** Further research is very unlikely to change our confidence in the estimate of effect.
- **Moderate quality:** Further research is likely to have an important impact on our confidence in the estimate of effect and may change the estimate.
- **Low quality:** Further research is very likely to have an important impact on our confidence in the estimate of effect and is likely to change the estimate.
- **Very low quality:** We are very uncertain about the estimate.

Evidence of effectiveness of counseling or other interventions to affect patients' choice of method

The review did not grade the *overall* body of evidence. However, the quality of individual studies was graded in accordance with USPSTF methodologies for doing so, i.e., Level I, Level II-1, Level II-2, Level II-3, Level III.

1a.7.3. Provide all other grades and associated definitions for strength of the evidence in the grading system.

Not applicable

1a.7.4. What is the time period covered by the body of evidence? (*provide the date range, e.g., 1990-2010*). Date range

Trussell (2011): 1958-2010
Mansour (2010): January 1990 to February 2008
Lopez (2015): through April 1, 2015
Zapata (2015): 1985-February 2011 with supplemental searches through 2014

QUANTITY AND QUALITY OF BODY OF EVIDENCE

1a.7.5. How many and what type of study designs are included in the body of evidence? (*e.g., 3 randomized controlled trials and 1 observational study*)

Evidence of contraceptive effectiveness & its impact on unintended pregnancy

- Trussell et al 2011: The review comprised results from clinical trials and surveys; the most recent review listed more than 350 studies, of which the majority was randomized controlled trials (Trussell 2011a).

- Mansour et al 2010: The authors identified and extracted information from 139 publications. Of the included studies, 47 assessed combined oral contraceptives (COCs), one assessed progestogen-only pills (POPs), three assessed the patch, three assessed the vaginal ring, 15 assessed implants, 16 assessed injectables, 31 assessed copper intrauterine devices (Cu-IUDs), nine assessed the levonorgestrel-releasing intrauterine system (LNGIUS), three assessed the male condom, four assessed other barrier methods, 11 assessed natural methods, and four assessed female sterilization. Overall, there were 64 publications of randomized controlled studies included in this review. A detailed description of each publication can be accessed from www.informahealthcare.com/doi/pdf/10.3109/13625180903427675.

- Lopez (2015). Fifteen RCTs were identified, with seven studies reported from 2010-2014.

Evidence of effectiveness of counseling or other interventions to affect patients' choice of method

- Zapata et al (2015): 22 studies (from 23 articles) met the inclusion criteria; 8 studies included use of more effective methods as an outcome. Seven of the 8 studies were randomized controlled trials, while the eighth utilized a pre-posttest study design.

1a.7.6. What is the overall quality of evidence across studies in the body of evidence? (*discuss the certainty or confidence in the estimates of effect particularly in relation to study factors such as design flaws, imprecision due to small numbers, indirectness of studies to the measure focus or target population*)

The quality of evidence is not described in either the Trussell (2011) or the Mansour (2010) publications. However, both reviews are substantially comprised of randomized controlled trials. The Lopez (2015) review determined that the overall body of evidence (comprised of 15 RCTs) was of moderate quality.

In Zapata et al (2011), 7 of the 8 studies were graded Level I (properly designed randomized controlled trial), and the 8th study was graded Level II-3 (evidence obtained from time series, uncontrolled trial).

ESTIMATES OF BENEFIT AND CONSISTENCY ACROSS STUDIES IN BODY OF EVIDENCE

1a.7.7. What are the estimates of benefit—magnitude and direction of effect on outcome(s) across studies in the body of evidence? (*e.g., ranges of percentages or odds ratios for improvement/ decline across studies, results of meta-analysis, and statistical significance*)

Evidence of contraceptive effectiveness & its impact on unintended pregnancy

- Trussell et al 2011: The key findings of this review are estimated failure rates for a wide range of contraceptive methods under "perfect" and "typical" use. The most recent findings – published in 2011 -- are that the most effective methods (LARC and sterilization) have a failure rate that is less than 1% per year under typical use; the moderately effective methods (shot, PPR, diaphragm) have a typical failure rate of 6-12% per year; the least effective methods have a typical failure rate of 18-28%; and not using any method at all has a failure rate of 85%.

- Mansour et al 2010: "Information was identified and extracted from 139 studies. One-year Pearl Indices reported for short-acting user-dependent hormonal methods were generally less than 2.5. Gross life-table rates for long-acting hormonal methods (implants and the levonorgestrel releasing-intrauterine system [LNG-IUS]) generally ranged between 0–0.6 per 100 at one year, but wider ranges (0.1–1.5 per 100) were observed for the copper intrauterine devices (0.1–1.4 per 100 for Cu-IUDs with surface area _300 mm2 and 0.6–1.5 per 100 for those with surface area5300 mm2). Barrier and natural methods were the least effective." The authors conclude that "the review broadly confirmed the hierarchy of contraceptive effectiveness in descending order as: (1) female sterilisation, long-acting hormonal contraceptives (LNG-IUS and implants); (2) Cu-IUDs with_300 mm2 surface area; (3) Cu-IUDs with5300 mm2 surface area and short-acting hormonal contraceptives (injectables, oral contraceptives, the patch and vaginal ring), and (4) barrier methods and natural methods."

- Lopez (2015): A meta-analysis showed that IUC use at six months was more likely with immediate insertion than with standard insertion (OR 2.04; 95% CI 1.10 to 4.09; participants=243; studies=4). Expulsion was more likely for the immediate group, but the confidence interval was wide (OR 4.89; 95% CI 1.47 to 16.32; participants =210; studies=4). The review concludes that the "benefit of effective contraception immediately after delivery may outweigh the disadvantage of increased risk for expulsion. Frequent prenatal visits during the third trimester provide the opportunity to discuss effective contraceptive methods and desired timing for initiation. Clinical follow-up can help detect early expulsion, as can educating women about expulsion signs and symptoms."

Evidence of effectiveness of counseling or other interventions to affect patients' choice of method

- Zapata (2015): Five of the 8 studies that examined use of more effective methods found an increased rate of use in the intervention vs control/comparison conditions. Three studies found no significant impact. No studies found a decreased rate of use of more effective contraceptive methods.

1a.7.8. What harms were studied and how do they affect the net benefit (benefits over harms)?

The harms were not noted in the cited reviews. However, CDC clinical recommendations on contraceptive safety explicitly address this question. CDC's "US Medical Eligibility Criteria for Contraceptive Use" (USMEC) describe what contraceptive methods are safe for women with a range of characteristics (e.g., age, postpartum) and medical conditions (e.g., infectious or chronic diseases). The citation for the USMEC recommendations is:

> CDC (2010). US Medical Eligibility Criteria for Contraceptive Use, MMWR Recommendations and Reports, 59 (RR04):1–85. Available online at: http://www.cdc.gov/reproductivehealth/UnintendedPregnancy/USMEC.htm."

The evidence on which the USMEC recommendations are based has been summarized in the following journal supplement:

> Contraception, Volume 82, Issue 1, Pages 1-118 (July 2010). Available online at: http://www.sciencedirect.com/science/journal/00107824/82/1

UPDATE TO THE SYSTEMATIC REVIEW(S) OF THE BODY OF EVIDENCE

1a.7.9. If new studies have been conducted since the systematic review of the body of evidence, provide for each new study: 1) citation, 2) description, 3) results, 4) impact on conclusions of systematic review.

Results from two large studies have been recently published, which provide additional evidence that: (a) long-acting reversible methods of contraception (LARC) are associated with reduced risk of unintended pregnancy, and (b) that the type of counseling provided is associated with selection of LARC methods by the client. The first study is a cluster-randomized trial led by researchers at the University of California – San Francisco (Harper 2015) and the second is a prospective cohort study that is known as "Project CHOICE" (Winner 2012).

UCSF trial (Harper et al 2015)

A cluster randomized trial was conducted in 2011-2013 to assess the effects of an intervention to increase patients' access to long-acting reversible contraceptives (LARCs) on pregnancy rates. A total of 40 clinics participated: 20 clinics were randomly assigned to receive evidence-based training on providing counselling and insertion of intrauterine devices (IUDs) or progestin implants, and 20 to provide standard care. Usual costs for contraception were maintained at all sites. Women aged 18-25 years attending family planning or abortion care visits and not desiring pregnancy in the next 12 months were recruited. The primary outcome was selection of an IUD or implant at the clinic visit and secondary outcome was pregnancy within 12 months. Generalised estimating equations for clustered data were used to measure the intervention effect on contraceptive selection, and survival analysis was used to assess pregnancy rates. Of 1500 women enrolled, more at intervention than control sites reported receiving counselling on IUDs or implants (565 [71%] of 797 vs 271 [39%] of 693, odds ratio 3·8, 95% CI 2·8-5·2) and more selected LARCs during the clinic visit (224 [28%] vs 117 [17%], 1·9, 1·3-2·8). The pregnancy rate was lower in intervention group than in the control group after family planning visits (7·9 vs 15·4 per 100 person-years), but not after abortion visits (26·5 vs 22·3 per 100 person-years). We found a significant intervention effect on pregnancy rates in women attending family planning visits (hazard ratio 0·54, 95% CI 0·34-0·85).

- Harper C, Rocca CH, Thompson KM, Morfesis J, Goodman S, Darney PD, Westhoff CL, Speidel JJ (2015). Reductions in pregnancy rates in the USA with long-acting reversible contraception: a cluster randomised trial. Lancet. Volume

386, No. 9993, p562–568, 8 August 2015

Project CHOICE (Secura et al 2014, Winner et al 2015)

The Contraceptive CHOICE Project was a prospective cohort study involving 9256 St. Louis area adolescent and adult women 14 to 45 years of age, in which women were counseled about the use of LARC methods to prevent unintended pregnancy. Participants were educated about reversible contraception, with an emphasis on the benefits of LARC methods, were provided with their choice of reversible contraception at no cost, and were followed for 2 to 3 years. Almost three-quarters of enrolled participants chose a LARC method when they were counseled about effectiveness and offered their choice of method at no charge, and continuation rates were high 2 years (77% for LARC users vs 41% for non-LARC users) and 3 years (67% for LARC users vs 31% for non-LARC users) after insertion. The contraceptive failure rate among participants using pills, patch, or ring was 4.55 per 100 participant-years, as compared with 0.27 among participants using long-acting reversible contraception (hazard ratio after adjustment for age, educational level, and history with respect to unintended pregnancy, 21.8; 95% confidence interval, 13.7 to 34.9).

- Winner B, Peipert J, Qiuhong Z, Buckel C, Madden T et al (2012). Effectiveness of Long-Acting Reversible Contraception, The New England Journal of Medicine, 366 (21): 1998-2007
- Diedrich, J.T., et al., *Three-year continuation of reversible contraception.* Am J Obstet Gynecol, 2015. **213**(5): p. 662 e1-8.
- O'Neil-Callahan, M., et al., *Twenty-four-month continuation of reversible contraception.* Obstet Gynecol, 2013. **122**(5): p. 1083-91.

_____

1a.8 OTHER SOURCE OF EVIDENCE – not applicable

If source of evidence is NOT from a clinical practice guideline, USPSTF, or systematic review, please describe the evidence on which you are basing the performance measure.

1a.8.1 What process was used to identify the evidence?

1a.8.2. Provide the citation and summary for each piece of evidence.

1. Evidence, Performance Gap, Priority – Importance to Measure and Report

Extent to which the specific measure focus is evidence-based, important to making significant gains in healthcare quality, and improving health outcomes for a specific high-priority (high-impact) aspect of healthcare where there is variation in or overall less-than-optimal performance. ***Measures must be judged to meet all subcriteria to pass this criterion and be evaluated against the remaining criteria.***

1a. Evidence to Support the Measure Focus – See attached Evidence Submission Form
nqf_evidence_Contraceptive_Care_Postpartum.docx

1b. Performance Gap

Demonstration of quality problems and opportunity for improvement, i.e., data demonstrating:

- considerable variation, or overall less-than-optimal performance, in the quality of care across providers; and/or
- disparities in care across population groups.

1b.1. **Briefly explain the rationale for this measure** *(e.g., the benefits or improvements in quality envisioned by use of this measure)*
Contraception is a highly effective clinical preventive service intended to prevent teen and unintended pregnancy and space births. The type of contraceptive method used by a woman is strongly associated with her risk of unintended pregnancy [1, 2]. The most effective methods (sterilization and the long-acting reversible methods of IUDs and implants) have a failure rate that is less than 1% per year under typical use; the moderately effective methods (shot, pill, patch, ring and diaphragm) have a typical failure rate of 6-12% per year; the least effective methods have a typical failure rate of 18-28%; and 85% of women will become pregnant in a year if they do not use any contraceptive method at all [2]. The proposed performance measure is based on the fact that some contraceptive methods are more effective than others in preventing unintended pregnancy, and hence, in spacing births among postpartum women.

The policy context is increasingly supportive of efforts to provide contraception in the postpartum period. ACOG, AAP and CDC recommend the provision of contraception in the postpartum period and describe how it can be done so safely (see evidence report for details). Financial barriers to immediate postpartum LARC insertion have been removed by numerous state Medicaid programs (i.e., by reimbursing separately for LARC in the immediate postpartum period, outside of the bundled delivery payment) [3]. The cost effectiveness of this practice has been documented [4, 5], and more states and private health plans are likely to adopt this policy in the near future. Given this context, we expect that use of the proposed clinical performance measures will encourage more providers to follow ACOG (2015) and CDC (2010, 2014) recommendations to:

- During prenatal visits, counsel all pregnant women about postpartum contraception, including the possibility of obtaining LARC and other methods of contraception at delivery, i.e., in the immediate postpartum period, and the effectiveness of the different methods as well as other factors that may help a woman decide the method that is best for her [6]. They will also take steps to make LARC available in the immediate postpartum inpatient setting [7].
- At the postpartum visit, counsel all women about contraception, which includes information about the effectiveness of the different methods as well as other factors that will help a woman choose the method that is best for her [6]. They will also take steps to make LARC available on a same-day basis in the outpatient setting [7].

References

1. Mansour, D., P. Inki, and K. Gemzell-Danielsson, Efficacy of contraceptive methods: a review of the literature. Eur J Contracept Reprod Health Care, 2010. 15 Suppl 2: p. S19-31.
2. Trussell, J., Contraceptive failure in the United States. Contraception, 2011. 83(5): p. 397-404.
3. Moniz, M.H., et al., Characterization of Medicaid policy for immediate postpartum contraception. Contraception, 2015. 92(6): p. 523-31.
4. Han, L., et al., Cost-effectiveness of immediate postpartum esonogestrel implant insertion for adolescent mothers Contraception. 86: p. 290–325.
5. Rodriguez, M.I., et al., Cost-benefit analysis of state- and hospital-funded postpartum intrauterine contraception at a university hospital for recent immigrants to the United States. Contraception, 2010. 81(4): p. 304-8.
6. CDC, Providing Quality Family Planning Services: Recommendations of the CDC and the U.S. Office of Population Affairs. MMWR Recommendations and Reports, 2014. 63(4): p. 1-54.

1b.2. **Provide performance scores on the measure as specified (current and over time) at the specified level of analysis.** *(This is*

required for endorsement maintenance. Include mean, std dev, min, max, interquartile range, scores by decile. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included). This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.

Performance scores for the measure are presented for two programs: (1) the Iowa Medicaid Enterprise, and (2) the Louisiana Medicaid Program. The Iowa claims dataset comprised 11,525 women who resided in one of six public health regions; the Louisiana dataset was comprised of 31,901 women ages 15-44 years who participated in one of 5 managed care organizations, and 33,980 women who resided in one of 9 public health regions.

Overall, the percentage of women that had been provided a most or moderately effective method of contraception by 3 days postpartum was 10.6% in Iowa and 9.7% in Louisiana. By day 60, the rates increased, ranging from 41.6 to 42.6% in Iowa, and from 42.7 to 45.3% in Louisiana. Performance varied by region and health plan. For example, the rates across Iowa's public regions ranged from 35 to 48%. In Louisiana, the rates across health plans ranged from 36 to 47%. (More details are provided in the testing report). It is desirable to have a very high proportion of postpartum women use a most or moderately effective method of contraception (e.g., more than 85%). For example, since approximately 41% of adolescent and adult Medicaid enrollees in Iowa were provided a most or moderately effective method of contraception at 60 days postpartum, there is potentially a 59 percentage point opportunity for improvement.

Overall, the percentage of women who were provided a LARC method by 3 days postpartum was less than 1% in both Iowa and Louisiana. By 60 days postpartum, LARC rates had increased to 10.7% in Iowa and 9.7% in Louisiana. (More details are provided in the evidence report). All health plans and regions in both states had LARC rates that were greater than 2%. The primary intent of the LARC measure is to identify populations in which LARC provision is noticeably low so that health programs can determine if there are barriers to access (e.g., less than 2%, or by calculating the median or mean and then identifying the reporting units where the rates of LARC provision are well below the median or mean). OPA maintains that the LARC measure should be used only to monitor access; and further, that it could be harmful to set a high benchmark for this measure, because doing so may incentivize coercive practices (Dehlendorf 2015, Gold 2014). Given this, the 3-day LARC rates found in Iowa and Louisian suggest that there may be barriers to provision in the immediate postpartum period; this result is not surprising given that Iowa and Louisiana Medicaid only started reimbursing for this service in 2014. However, if low rates persist once the policy has been in place for a longer period of time, steps might need to be taken to remove additional barriers. The results for LARC at 60 days postpartum show that performance increased notably; the fact that no public health regions have a rate of less than 2 percent suggests that there is some access to LARC in the state overall, in each region and in each health plan.

Some have raised questions about how to interpret the measure´s rates given the protective effects of breastfeeding. Although the Lactational Amenorrhea Method (LAM) is a highly effective, temporary method of contraception that can be used in the postpartum period, a premise underpinning these measures is that providers can encourage the provision of contraception in the postpartum period while simultaneously encouraging breastfeeding. CDC and ACOG recommendations state that a wide range of contraceptive methods can be used safely by women who are breastfeeding including: progestin-only pill, shot or implant; IUDs; and sterilization [CDC 2010, ACOG 2015]. LAM is more than 98% effective at preventing pregnancy under perfect use [Trussell 2011], but it can be challenging to implement because it requires that all of the following three conditions must be met: (1) the mother´s monthly bleeding has not returned; (2) the baby is exclusively breastfed and is fed often, day and night; and (3) the baby is less than 6 months old [WHO 2011]. In the US, rates of exclusive breastfeeding (not necessarily LAM) drop quickly in the postpartum period. For example, national surveillance data from CDC show that although 80% of U.S. children were ever breastfed, only 43% were exclusively breastfed at 3 months postpartum, and 22% were exclusively breastfed at six months [CDC/DNPAO].

Many women may find it difficult to return for contraceptive services after the 6-week postpartum visit. By not missing the opportunity to provide contraception in the immediate postpartum period, or the 6-week postpartum visit, providers will help women who use LAM to transition to another method of contraception without any interruption in pregnancy prevention.

References

- ACOG Long-acting reversible contraception: Implants and intrauterine devices, in Practice Bulletin. 2015 (reaffirmed), American College of Obstetricians and Gynecologists: Washington, DC. p. 1-13.
- CDC, U.S. Medical Eligibility Criteria for Contraceptive Use 2010. MMWR Recomm Rep, 2010. 59(RR-4): p. 1-86.
- CDC/Division of Nutrition, P.A.a.O. Breastfeeding among U.S. Children Born 2002–2012, CDC National Immunization Surveys [cited 2016 January 30]; Available from: http://www.cdc.gov/breastfeeding/data/nis_data/index.htm.
- Trussell, J., Contraceptive failure in the United States. Contraception, 2011. 83(5): p. 397-404.
- WHO/Department of Reproductive Health and Research & Johns Hopkins Bloomberg School of Public Health/Center for

Communication Programs, Family Planning: A Global Handbook for Providers. 2011, Baltimore and Geneva: World Health Organization & Johns Hopkins University.

1b.3. **If no or limited performance data on the measure as specified is reported in 1b2, then provide a summary of data from the literature that indicates opportunity for improvement or overall less than optimal performance on the specific focus of measurement.**
A recent analysis of data from CDC's Pregnancy Risk Assessment Monitoring System, 2011-2012 (PRAMS) showed that in the 2-6 month postpartum period only 50% of women reported using a most or moderately effective method of contraception, and 13.6% were using a LARC method (CDC, unpublished data – see testing report).

1b.4. **Provide disparities data from the measure as specified (current and over time) by population group, e.g., by race/ethnicity, gender, age, insurance status, socioeconomic status, and/or disability.** *(This is required for endorsement maintenance. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities include.) This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*

1b.5. **If no or limited data on disparities from the measure as specified is reported in 1b4, then provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement. Include citations.**
An analysis of data from the Pregnancy Risk Assessment Monitoring System (PRAMS), 2011-2012, was conducted to explore disparities in the use of most and moderately effective and LARC methods of contraception (see table of PRAMS results in the testing report) (unpublished data, CDC). This analysis suggests that there are statistically significant differences by age, marital status and some income categories for use of most and moderately effective methods and LARC methods. However, there were no significant differences by race/ethnicity, and most income categories.

1c. High Priority (previously referred to as High Impact)

The measure addresses:

- a specific national health goal/priority identified by DHHS or the National Priorities Partnership convened by NQF; OR
- a demonstrated high-priority (high-impact) aspect of healthcare (e.g., affects large numbers of patients and/or has a substantial impact for a smaller population; leading cause of morbidity/mortality; high resource use (current and/or future); severity of illness; and severity of patient/societal consequences of poor quality).

1c.1. **Demonstrated high priority aspect of healthcare**
Affects large numbers, A leading cause of morbidity/mortality, Frequently performed procedure, High resource use
1c.2. **If Other:**

1c.3. **Provide epidemiologic or resource use data that demonstrates the measure addresses a high priority aspect of healthcare. List citations in 1c.4.**
Over the course of a lifetime, most individuals will make decisions related to childbearing, i.e., how to prevent or achieve pregnancy so that they can attain their desired number and spacing of children. Of the 310 million people in the United States, 62 million (20%) are women of reproductive age, 15-44 years; approximately 20% are men of that same age [1]. Thirty-eight million women are in need of contraceptive services because they are at risk for unintended pregnancy – that is, they are sexually active, are able to get pregnant and want to avoid or space pregnancy [2]. Contraception is highly effective at preventing unintended pregnancy; and cost-effective, with $6-8 saved for every $1 invested [3-10].

Yet use of the most and moderately effective methods of contraception remain relatively low in women overall -- including in the postpartum period. A recent analysis of data from CDC's Pregnancy Risk Assessment Monitoring System, 2011-2012 (PRAMS) showed that in the 2-6 month postpartum period only 50% of women reported using a most or moderately effective method of contraception, and 13.6% were using a LARC method (CDC, unpublished data – see evidence report). Due to these patterns of postpartum contraceptive use, one-third of pregnancies are spaced within 18 months (i.e., they do not follow recommendations contained in Healthy People 2020 objectives) [11, 12]. Further, the rate of unintended pregnancy is high in the United States with more than one-half (51%) of the 6.7 million pregnancies each year (3.2 million) being unintended [13]. Finally, despite recent reductions, each year more than 58 of every 1000 women aged 15-19 years become pregnant [14], and 250,000 adolescents give birth [15].

The consequences of closely spaced births, unintended and teen pregnancy are wide-ranging. Appropriate interpregnancy intervals have been associated with improved infant health outcomes, including a reduction in the rate of infants who are born preterm and/or low birth weight [16-27]. Approximately 1 out of every 8 pregnancies in the United States results in preterm birth, and infant mortality rates remain high relative to other developed countries [15, 28, 29]. Taxpayers pay a high price for the nation's high rate of poorly spaced births, teen and unintended pregnancy; for example, the direct medical cost of publicly funded births was estimated at $21 billion per year (this figure includes costs for prenatal care, labor and delivery, post-partum care, and one year of infant care) [31]. The cost of teen pregnancy alone has been estimated at $9.4 billion per year [30]; and adolescents who give birth as teens are more likely to achieve less education and lower incomes while their children may experience higher rates of negative outcomes such as poorer health, lowered academic achievement, and higher rates of teen pregnancy for female children and incarceration for male children [32].

The prevention of teen and unintended pregnancy and improved rates of birth spacing have repeatedly been identified as national priorities. Most recently, in 2015 the Institute of Medicine (IOM) recognized the importance of unintended pregnancy when they included it as one of 15 core measures that constitute the most vital signs for the nation's health and health care [33]. The United States' National Prevention Strategy [34] and Healthy People 2020 Objectives also include several specific objectives focused on appropriate inter-pregnancy intervals, unintended pregnancy and use of contraception (see www.healthypeople.gov/2020/topics-objectives/topic/family-planning).

1c.4. Citations for data demonstrating high priority provided in 1a.3

1. US Census Bureau. Age and Sex Composition in the United States: 2011. 2013 [cited 2013 October 17]; Available from: http://www.census.gov/population/age/data/2011comp.html.
2. Frost, J. and et al, Contraceptive Needs and Services 2013 Update. 2015, Guttmacher Institute: New York, NY.
3. Cleland, K., et al., Family planning as a cost-saving preventive health service. N Engl J Med 2011. 364: p. e37.
4. Foster, D.G., et al., Cost-savings from the provision of specific contraceptive methods in 2009. Womens Health Issues, 2013. 23(4): p. e265-71.
5. Frost, J., L. Finer, and A. Tapales, The Impact of Publicly Funded Family Planning Clinic Services on Unintended Pregnancies and Government Cost Savings. Journal of Health Care for the Poor and Underserved, 2008. 19(3): p. 778-796.
6. Thomas, A., Three Strategies to Prevent Unintended Pregnancy. Journal of Policy Analysis and Management 2012. 31(2): p. 280-311.
7. Trussell, J., Update on and correction to the cost-effectiveness of contraceptives in the United States. Contraception, 2012. 85(6): p. 611.
8. Trussell, J., et al., Cost-effectiveness analysis of levonorgestrel-releasing intrauterine system (LNG-IUS) 13.5 mg in contraception. Contraception, 2014. 89(5): p. 451-9.
9. Trussell, J., et al., Achieving cost-neutrality with long-acting reversible contraceptive methods. Contraception, 2015. 91(1): p. 49-56.
10. Trussell, J., et al., Burden of unintended pregnancy in the United States: potential savings with increased use of long-acting reversible contraception. Contraception, 2013. 87(2): p. 154-61.
11. Copen, C., M. Thomas, and S. Kirmeyer, Interpregnancy intervals in the United States: Data from the birth certificate and the National Survey of Family Growth. , in National vital statistics reports. 2015, National Center for Health Statistics: Hyattsville, MD.
12. Gemmill, A. and L.D. Lindberg, Short interpregnancy intervals in the United States. Obstet Gynecol, 2013. 122(1): p. 64-71.
13. Finer, L.B. and M.R. Zolna, Shifts in intended and unintended pregnancies in the United States, 2001-2008. Am J Public Health, 2014. 104 Suppl 1: p. S43-8.
14. Curtin, S.C., et al., Pregnancy rates for U.S. women continue to drop. NCHS Data Brief, 2013(136): p. 1-8.
15. Hamilton, B., et al., Births: Final data for 2014, N.C.f.H. Statistics, Editor. 2015, Centers for Disease Control and Prevention: Hyattsville, MD.
16. Conde-Agudelo, A., et al., Effects of birth spacing on maternal, perinatal, infant, and child health: a systematic review of causal mechanisms. Stud Fam Plann, 2012. 43(2): p. 93-114.
17. Conde-Agudelo, A., A. Rosas-Bermudez, and A.C. Kafury-Goeta, Birth spacing and risk of adverse perinatal outcomes: a meta-analysis. JAMA, 2006. 295(15): p. 1809-23.
18. Conde-Agudelo, A., A. Rosas-Bermudez, and A.C. Kafury-Goeta, Effects of birth spacing on maternal health: a systematic review. Am J Obstet Gynecol, 2007. 196(4): p. 297-308.
19. DeFranco, E.A., et al., Influence of interpregnancy interval on neonatal morbidity. Am J Obstet Gynecol, 2015. 212(3): p. 386.e1-9.
20. Gipson, J.D., M.A. Koenig, and M.J. Hindin, The effects of unintended pregnancy on infant, child, and parental health: a review of the literature. Stud Fam Plann, 2008. 39(1): p. 18-38.

21. Salihu, H.M., et al., The impact of birth spacing on subsequent feto-infant outcomes among community enrollees of a federal healthy start project. J Community Health, 2012. 37(1): p. 137-42.
22. Sawhill, I., Q. Karpilow, and J. Venator, The impact of unintended childbearing on future generations, C.o.C.a. Families, Editor. 2014, Brookings Institution: Washington, DC.
23. Shah, P.S., et al., Intention to become pregnant and low birth weight and preterm birth: a systematic review. Matern Child Health J, 2011. 15(2): p. 205-16.
24. Tsui, A.O., R. McDonald-Mosley, and A.E. Burke, Family planning and the burden of unintended pregnancies. Epidemiol Rev, 2010. 32(1): p. 152-74.
25. Zhu, B.P., et al., Effect of the interval between pregnancies on perinatal outcomes. N Engl J Med, 1999. 340(8): p. 589-94.
26. Crowne, S.S., et al., Relationship between birth spacing, child maltreatment, and child behavior and development outcomes among at-risk families. Matern Child Health J, 2012. 16(7): p. 1413-20.
27. Nerlander, L.M., et al., Short interpregnancy interval associated with preterm birth in U S adolescents. Matern Child Health J, 2015. 19(4): p. 850-8.
28. MacDorman, M.F., Race and ethnic disparities in fetal mortality, preterm birth, and infant mortality in the United States: an overview. Semin Perinatol, 2011. 35(4): p. 200-8.
29. MacDorman, M.F. and T.J. Mathews, BirthStats: percentage of preterm births, United States and selected European countries, 2004. Birth, 2010. 37(2): p. 168.
30. Hoffman, S., Counting It Up: The Public Costs of Teen Childbearing. 2010, The National Campaign to Prevent Teen and Unplanned Pregnancy: Washington, DC.
31. Sonfield, A. and K. Kost, Public Costs from Unintended Pregnancies and the Role of Public Insurance Programs in Paying for Pregnancy and Infant Care: Estimates for 2010, T.G. Institute, Editor. 2015, The Guttmacher Institute: New York, NY.
32. Hoffman, S. and R. Maynard, Kids having kids: Economic costs and social consequences of teen pregnancy 2nd ed. 2008, Washington, DC: Urban Institute Press
33. Institute of Medicine, C.o.C.M.f.B.H.a.L.C., Vital Signs: Core Metrics for Health and Health Care Progress, M.E. Blumenthal D, McGinnis M, Editors, Editor. 2015, National Academies Press: Washington DC.
34. National Prevention Council, National Prevention Strategy. 2011, U.S. Department of Health and Human Services, Office of the Surgeon General, : Washington DC.

1c.5. If a PRO-PM (e.g. HRQoL/functional status, symptom/burden, experience with care, health-related behaviors), provide evidence that the target population values the measured PRO and finds it meaningful. *(Describe how and from whom their input was obtained.)*

2. Reliability and Validity—Scientific Acceptability of Measure Properties

Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. ***Measures must be judged to meet the subcriteria for both reliability and validity to pass this criterion and be evaluated against the remaining criteria.***

2a.1. Specifications The measure is well defined and precisely specified so it can be implemented consistently within and across organizations and allows for comparability. eMeasures should be specified in the Health Quality Measures Format (HQMF) and the Quality Data Model (QDM).

De.5. **Subject/Topic Area** *(check all the areas that apply):*
Perinatal and Reproductive Health : Newborn, Perinatal and Reproductive Health : Perinatal, Prevention

De.6. **Cross Cutting Areas** *(check all the areas that apply):*
Access, Prevention

S.1. **Measure-specific Web Page** *(Provide a URL link to a web page specific for this measure that contains current detailed specifications including code lists, risk model details, and supplemental materials. Do not enter a URL linking to a home page or to general information.)*

S.2a. **If this is an eMeasure**, HQMF specifications must be attached. Attach the zipped output from the eMeasure authoring tool (MAT) - if the MAT was not used, contact staff. (Use the specification fields in this online form for the plain-language description of the specifications)
This is not an eMeasure **Attachment:**

S.2b. **Data Dictionary, Code Table, or Value Sets** *(and risk model codes and coefficients when applicable) must be attached. (Excel or csv file in the suggested format preferred - if not, contact staff)*
Attachment **Attachment:** Codes_2014_and_2015_Postpartum_Contraception.xlsx

S.3. **For endorsement maintenance**, please briefly describe any changes to the measure specifications since last endorsement date and explain the reasons.

S.4. **Numerator Statement** *(Brief, narrative description of the measure focus or what is being measured about the target population, i.e., cases from the target population with the target process, condition, event, or outcome)*
IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
Primary measure: Women ages 15 through 44 who had a live birth and were provided a most (sterilization, intrauterine device, implant) or moderately (pill, patch, ring, injectable, diaphragm) effective method of contraception within 3 and 60 days of delivery.
Sub-measure: Women ages 15 through 44 who had a live birth and were provided a long-acting reversible method of contraception (LARC) within 3 and 60 days of delivery.

S.5. **Time Period for Data** *(What is the time period in which data will be aggregated for the measure, e.g., 12 mo, 3 years, look back to August for flu vaccination? Note if there are different time periods for the numerator and denominator.)*
A twelve-month period of time is used.

S.6. **Numerator Details** *(All information required to identify and calculate the cases from the target population with the target process, condition, event, or outcome such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
IF an OUTCOME MEASURE, describe how the observed outcome is identified/counted. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
The target population is women ages 15- 44 who had a live birth and were provided a most or moderately effective method (primary measure) or a LARC method (sub-measure) of contraception. All claims codes are found in the attached Excel files. To identify the numerator, follow these steps:

Step 1 Use the codes in Table PCU-C to identify women who were provided a most (sterilization, IUD, implant) or moderately (injection, oral pills, patch, ring, or diaphragm) effective method of contraception in the measurement year. Use the codes in PCU-E to identify women who were provided a LARC method.

Step 2 The long-acting reversible contraceptive (LARC) methods of intrauterine devices (IUD) and implants can be removed at the woman's request so adjustments must be made to reflect this. Use the codes in Table PCU -D to identify women who had their IUD or implant removed at any point during the measurement year. Check to see if they had an IUD or implant reinserted on the same or a subsequent date.

[For the primary measure] If there is no code indicating reinsertion, use the codes in Table PCU -E to determine whether a woman was provided another most or moderately effective method. Do so by looking back over the 30 days prior to the removal (since a woman may receive a prescription for another method prior to the removal) as well as the period after the LARC removal. If there is no code for reinsertion or provision of another most or moderately effective method, consider them as a non-user.

Step 3 Subtract the number of women identified as non-users of contraception in step 2 from those identified in step 1 to determine the numerator. Calculate the numerator separately for the two age groups: adolescents and adults.

S.7. **Denominator Statement** *(Brief, narrative description of the target population being measured)*
Women ages 15 through 44 who had a live birth in a 12-month measurement year.

S.8. **Target Population Category** *(Check all the populations for which the measure is specified and tested if any):*
Children's Health, Maternal Health

S.9. **Denominator Details** *(All information required to identify and calculate the target population/denominator such as definitions, specific data collection items/responses , code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
The target population is women ages 15 through 44 who had a live birth in a 12-month measurement year. In a Medicaid population, this includes women who were enrolled from the date of delivery to 60 days postpartum.

S.10. **Denominator Exclusions** *(Brief narrative description of exclusions from the target population)*
The following categories are excluded from the denominator: (1) deliveries that did not end in a live birth (i.e., miscarriage, ectopic, stillbirth or induced abortion); and (2) deliveries that occurred during the last two months of the measurement year.

S.11. **Denominator Exclusion Details** *(All information required to identify and calculate exclusions from the denominator such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
Women are excluded from the denominator if they did not have an opportunity to receive contraception in the postpartum period (defined as within 60 days of delivery). All claims codes are found in the attached Excel files. Follow the steps below to identify the eligible population:

Step 1 Identify live births and deliveries by using codes in Table PCU-A (we used the codes developed for the HEDIS measure of Prenatal and Postnatal care). Some women may have more than one delivery in the measurement year; the measure is designed to identify unique live births (defined as those that occur >180 days apart) rather than women who had a live birth.

Step 2 Exclude deliveries that did not end in a live birth (i.e., miscarriage, ectopic, stillbirth, or pregnancy termination) by using the codes in Table PCU-B . We used the codes developed to identify live births for the HEDIS measure of Prenatal and Postnatal Care.

Step 3 Exclude deliveries that occurred during the last 2 months of the measurement year. These deliveries should be excluded from the denominator because there may not have been an opportunity to provide the mother with contraception during the postpartum period. A two-month period was selected because the American College of Obstetricians and Gynecologists (ACOG) recommends having a postpartum visit by 6 weeks, and an additional 2 weeks was added to allow for reasonable delays in attending the postpartum visit.

S.12. **Stratification Details/Variables** *(All information required to stratify the measure results including the stratification variables,*

definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format with at S.2b)
The primary stratification variable is age, so that adolescents can be examined separately from adult women. We propose this stratification for purposes of QI but not as a method of risk adjustment. Teen pregnancy is worthy of a separate focus because of the large potential negative impact on the life of the teen and her child(ren), and the existence of unique programs and contraceptive counseling approaches tailored to this population. In the pilot data presented, we used age groups that are consistent with Center for Medicaid and CHIP Servies (CMCS) reporting requirements, i.e., adolescents are defined as 15-20 years and adults are 21-44 years of age.

S.13. Risk Adjustment Type (Select type. Provide specifications for risk stratification in S.12 and for statistical model in S.14-15)
No risk adjustment or risk stratification
If other:

S.14. Identify the statistical risk model method and variables *(Name the statistical method - e.g., logistic regression and list all the risk factor variables. Note - risk model development and testing should be addressed with measure testing under Scientific Acceptability)*
We do not believe that risk adjustment is justified. Although there are possible variations in contraceptive provision by socio-demographic characteristics, the reason for those patterns is based on modifiable clinical and programmatic considerations rather than differing biological responses to contraception. More detailed information about variations in provision of most and moderately effective contraceptive methods by socio-demographic characteristics of women of reproductive age, can be found in the testing report.

S.15. Detailed risk model specifications *(must be in attached data dictionary/code list Excel or csv file. Also indicate if available at measure-specific URL identified in S.1.)*
Note: Risk model details (including coefficients, equations, codes with descriptors, definitions), should be provided on a separate worksheet in the suggested format in the Excel or csv file with data dictionary/code lists at S.2b.

S.15a. Detailed risk model specifications (*if not provided in excel or csv file at S.2b*)

S.16. Type of score:
Rate/proportion
If other:

S.17. Interpretation of Score *(Classifies interpretation of score according to whether better quality is associated with a higher score, a lower score, a score falling within a defined interval, or a passing score)*
Better quality = Score within a defined interval

S.18. Calculation Algorithm/Measure Logic *(Describe the calculation of the measure score as an ordered sequence of steps including identifying the target population; exclusions; cases meeting the target process, condition, event, or outcome; aggregating data; risk adjustment; etc.)*
Step 1 Identify live births that occurred in the measurement year. Some women may have more than one delivery in the measurement year; the measure is designed to identify unique live births (defined as those that occur >180 days apart) rather than women who had a live birth.

Step 2 Exclude the following deliveries:
- Those that did not end in a live birth (i.e., miscarriage, ectopic, stillbirth, or pregnancy termination).
- Those that occurred during the last 2 months of the measurement year. These deliveries should be excluded from the denominator because there may not have been an opportunity to provide the mother with contraception during the postpartum period.

Step 3 Define the numerator by identifying women who were provided a most (sterilization, IUD, implant) or moderately (injection, oral pills, patch, ring, or diaphragm) effective method of contraception in the measurement year (primary measure). For the sub-measure, identify women who were provided a LARC method.

Step 4 Determine the date that the contraceptive method was provided, to identify women who were rovided it: (a) within 3 day sof delivery, and (b) within 60 days of delivery.

Step 5 Divide the number of women using a most or moderately effective method [or LARC, for the sub-measure] by the number of eligible women in the denominator. Calculate the rates separately for the two age groups: adolescents and adults.

S.19. **Calculation Algorithm/Measure Logic Diagram URL or Attachment** *(You also may provide a diagram of the Calculation Algorithm/Measure Logic described above at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
Available in attached appendix at A.1

S.20. **Sampling** *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*
IF a PRO-PM, identify whether (and how) proxy responses are allowed.
Not applicable.

S.21. **Survey/Patient-reported data** *(If measure is based on a survey, provide instructions for conducting the survey and guidance on minimum response rate.)*
IF a PRO-PM, specify calculation of response rates to be reported with performance measure results.
Not applicable.

S.22. **Missing data** (specify how missing data are handled, e.g., imputation, delete case.)
Required for Composites and PRO-PMs.
Not applicable.

S.23. **Data Source** *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*
If other, please describe in S.24.
Administrative claims

S.24. **Data Source or Collection Instrument** *(Identify the specific data source/data collection instrument e.g. name of database, clinical registry, collection instrument, etc.)*
IF a PRO-PM, identify the specific PROM(s); and standard methods, modes, and languages of administration.
Adminisrative claims data are used to calculate the measure. The data request should include an eligibility file, paid and denied claims with diagnosis codes and procedures codes (HCPCS, CPT, and ICD-9-PCS/ICD-10-PCS), as well as NDC codes.

S.25. **Data Source or Collection Instrument** *(available at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*

S.26. **Level of Analysis** *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*
Health Plan, Population : Regional

S.27. **Care Setting** *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*

If other:

S.28. **COMPOSITE Performance Measure** - Additional Specifications *(Use this section as needed for aggregation and weighting rules, or calculation of individual performance measures if not individually endorsed.)*

2a. Reliability – See attached Measure Testing Submission Form
2b. Validity – See attached Measure Testing Submission Form
Codes_2014_and_2015_Postpartum_Contraception-635918259676623524.xlsx,nqf_testing_Contraceptive_Care_Postpartum.docx

NATIONAL QUALITY FORUM—Measure Testing (subcriteria 2a2, 2b2-2b7)

Measure Number (*if previously endorsed*)**:** Click here to enter NQF number

Measure Title: Contraceptive Care – POSTPARTUM

Date of Submission: Click here to enter a date

Type of Measure: Intermediate Clinical Outcome

☐ Composite – ***STOP – use composite testing form***	☒ Outcome (*including PRO-PM*)
☐ Cost/resource	☐ Process
☐ Efficiency	☐ Structure

Instructions

- Measures must be tested for all the data sources and levels of analyses that are specified. ***If there is more than one set of data specifications or more than one level of analysis, contact NQF staff*** about how to present all the testing information in one form.
- **For all measures, sections 1, 2a2, 2b2, 2b3, and 2b5 must be completed.**
- **For outcome and resource use measures**, section **2b4** also must be completed.
- If specified for **multiple data sources/sets of specificaitons** (e.g., claims and EHRs), section **2b6** also must be completed.
- Respond to all questions as instructed with answers immediately following the question. All information on testing to demonstrate meeting the subcriteria for reliability (2a2) and validity (2b2-2b6) must be in this form. An appendix for *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 20 pages (*incuding questions/instructions;* minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.
- For information on the most updated guidance on how to address sociodemographic variables and testing in this form refer to the release notes for version 6.6 of the Measure Testing Attachment.

Note: The information provided in this form is intended to aid the Steering Committee and other stakeholders in understanding to what degree the testing results for this measure meet NQF's evaluation criteria for testing.

2a2. Reliability testing [10] demonstrates the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise. For **PRO-PMs and composite performance measures**, reliability should be demonstrated for the computed performance score.

2b2. Validity testing [11] demonstrates that the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality. For **PRO-PMs and composite performance measures**, validity should be demonstrated for the computed performance score.

2b3. Exclusions are supported by the clinical evidence; otherwise, they are supported by evidence of sufficient frequency of occurrence so that results are distorted without the exclusion; [12]

AND
If patient preference (e.g., informed decision-making) is a basis for exclusion, there must be evidence that the exclusion impacts performance on the measure; in such cases, the measure must be specified so that the information about patient preference and the effect on the measure is transparent (e.g., numerator category computed separately, denominator exclusion category computed separately). [13]

2b4. For outcome measures and other measures when indicated (e.g., resource use):

- **an evidence-based risk-adjustment strategy** (e.g., risk models, risk stratification) is specified; is based on patient factors (including clinical and sociodemographic factors) that influence the measured outcome and are present at start of care; [14,15] and has demonstrated adequate discrimination and calibration

OR

- rationale/data support no risk adjustment/ stratification.

2b5. Data analysis of computed measure scores demonstrates that methods for scoring and analysis of the specified measure allow for **identification of statistically significant and practically/clinically meaningful** [16] **differences in performance**;
OR
there is evidence of overall less-than-optimal performance.

2b6. If multiple data sources/methods are specified, there is demonstration they produce comparable results.

2b7. For **eMeasures, composites, and PRO-PMs** (or other measures susceptible to missing data), analyses identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias.

Notes
10. Reliability testing applies to both the data elements and computed measure score. Examples of reliability testing for data elements include, but are not limited to: inter-rater/abstractor or intra-rater/abstractor studies; internal consistency for multi-item scales; test-retest for survey items. Reliability testing of the measure score addresses precision of measurement (e.g., signal-to-noise).
11. Validity testing applies to both the data elements and computed measure score. Validity testing of data elements typically analyzes agreement with another authoritative source of the same information. Examples of validity testing of the measure score include, but are not limited to: testing hypotheses that the measures scores indicate quality of care, e.g., measure scores are different for groups known to have differences in quality assessed by another valid quality measure or method; correlation of measure scores with another valid indicator of quality for the specific topic; or relationship to conceptually related measures (e.g., scores on process measures to scores on outcome measures). Face validity of the measure score as a quality indicator may be adequate if accomplished through a systematic and transparent process, by identified experts, and explicitly addresses whether performance scores resulting from the measure as specified can be used to distinguish good from poor quality.
12. Examples of evidence that an exclusion distorts measure results include, but are not limited to: frequency of occurrence, variability of exclusions across providers, and sensitivity analyses with and without the exclusion.
13. Patient preference is not a clinical exception to eligibility and can be influenced by provider interventions.
14. Risk factors that influence outcomes should not be specified as exclusions
15. With large enough sample sizes, small differences that are statistically significant may or may not be practically or clinically meaningful. The substantive question may be, for example, whether a statistically significant difference of one percentage point in the percentage of patients who received smoking cessation counseling (e.g., 74 percent v. 75 percent) is clinically meaningful; or whether a statistically significant difference of $25 in cost for an episode of care (e.g., $5,000 v. $5,025) is practically meaningful. Measures with overall less-than-optimal performance may not demonstrate much variability across providers.

1. DATA/SAMPLE USED FOR ALL TESTING OF THIS MEASURE

Often the same data are used for all aspects of measure testing. In an effort to eliminate duplication, the first five questions apply to all measure testing. If there are differences by aspect of testing,(e.g., reliability vs. validity) be sure to indicate the specific differences in question 1.7.

1.1. What type of data was used for testing? (*Check all the sources of data identified in the measure specifications and data used for testing the measure. Testing must be provided for all the sources of data specified and intended for measure implementation.* ***If different data sources are used for the numerator and denominator, indicate N [numerator] or D [denominator] after the checkbox.***)

Measure Specified to Use Data From: (***must be consistent with data sources entered in S.23***)	**Measure Tested with Data From:**
☐ abstracted from paper record	☐ abstracted from paper record
☒ administrative claims	☒ administrative claims
☐ clinical database/registry	☐ clinical database/registry
☐ abstracted from electronic health record	☐ abstracted from electronic health record
☐ eMeasure (HQMF) implemented in EHRs	☐ eMeasure (HQMF) implemented in EHRs
☐ other: Click here to describe	☐ other: Click here to describe

1.2. If an existing dataset was used, identify the specific dataset (*the dataset used for testing must be consistent with the measure specifications for target population and healthcare entities being measured; e.g., Medicare Part A claims, Medicaid claims, other commercial insurance, nursing home MDS, home health OASIS, clinical registry*).

Two claims datasets were used for testing:

(1) Data from the Iowa Medicaid Program (IME). The IME dataset comprised all female Medicaid clients aged 15-44 years who resided in 6 public health regions, and participated in either fee-for-service care or in two health plans. In 2013, Medicaid services in Iowa were provided primarily on a fee-for-service basis, although a small percentage of clients (approximately 10%) were provided care through one of two managed care organizations (MCO); for this reason, reliability was not assessed for this level using the IME data.

(2) Data from the Louisiana Medicaid program (LM). The LM dataset for 2014 included all female Medicaid enrollees aged 15-44 years who resided in any of the 9 Louisiana Public Health Regions, participated in either fee-for-service care or in one of the five health plans, and participated in either the general Medicaid program or the state's family planning waiver program. Louisiana Medicaid provides contraceptive services to women through its general Medicaid program and its family planning waiver program (Take Charge and Take Charge Plus). Take Charge Plus superseded Take Charge on September 1, 2014. Services are available to Louisiana uninsured Louisiana residents not eligible for Medicaid, Louisiana's CHIP program, or Medicare and who do not have private insurance. To be eligible for Take Charge services, the individual must be a woman between the ages of 19 and 44 with income at or below 200% percent of the federal poverty level. The guidelines for Take Charge Plus include women or men of any age with income at or below 138% of the federal poverty level. Take Charge benefits were extended until December 31, 2014, for those who, because of income, would not be eligible for Take Charge Plus. In 2014, Medicaid services in Louisiana (excluding Medicaid-Medicare dual-eligibles) were provided primarily by 5 Bayou Health managed care plans (3 traditional prepaid plans and 2 shared savings plans). Approximately 15% of the non dual-eligible Medicaid population was continuously enrolled in traditional fee-for-service Medicaid.

1.3. What are the dates of the data used in testing? January 1 2013 – December 31 2014

Data from IME covered the period January 1, 2013 – December 31, 2013.

Data from LM covered the period January 1 2014 – December 31 2014.

1.4. What levels of analysis were tested? (*testing must be provided for all the levels specified and intended for measure implementation, e.g., individual clinician, hospital, health plan*)

Measure Specified to Measure Performance of: (*must be consistent with levels entered in item S.26*)	**Measure Tested at Level of:**
☐ individual clinician	☐ individual clinician
☐ group/practice	☐ group/practice
☐ hospital/facility/agency	☐ hospital/facility/agency
☒ health plan	☒ health plan
☒ other: Public health region	☒ other: Public health region

1.5. How many and which measured entities were included in the testing and analysis (by level of analysis and data source)? (*identify the number and descriptive characteristics of measured entities included in the analysis (e.g., size, location, type); if a sample was used, describe how entities were selected for inclusion in the sample*)

Reliability

The measure was tested at two levels, as shown in the table below.

Level	Number of measured entities	Data Source
Health plan (Medicaid managed care)	n.a. 5	Iowa Medicaid Enterprise Louisiana Medicaid
Public health region	6 9	Iowa Medicaid Enterprise Louisiana Medicaid

Validity

A panel of experts assessed the measure's face validity.

1.6. How many and which patients were included in the testing and analysis (by level of analysis and data source)? (*identify the number and descriptive characteristics of patients included in the analysis (e.g., age, sex, race, diagnosis); if a sample was used, describe how patients were selected for inclusion in the sample*)

Level of analysis	Number of patients					
	Iowa Medicaid			Louisiana Medicaid		
	15 - 20	21 -44	15 - 44	15 - 20	21 -44	15 - 44
Health plan						
MCO1	NA	NA	NA	945	4151	5096
MCO2				941	4024	4965
MCO3				975	3840	4815
MCO4				1291	5749	7040
MCO5				1716	8269	9985
Total: Health Plan				**5868**	**26033**	**31901**
Public health region						
Region 1	606	2787	3393	926	5734	6660
Region 2	145	791	936	760	3880	4640
Region 3	314	1412	1726	540	2495	3035
Region 4	244	979	1223	941	3659	4600
Region 5	331	1183	1514	485	1765	2250
Region 6	503	2230	2733	501	1777	2278
Region 7	--	--	--	760	3162	3922
Region 8	--	--	--	568	2361	2929
Region 9	--	--	--	666	3000	3666
Total: Public Health Region	**2143**	**9382**	**11525**	**6147**	**27833**	**33980**

1.7. If there are differences in the data or sample used for different aspects of testing (e.g., reliability, validity, exclusions, risk adjustment), identify how the data or sample are different for each aspect of testing reported below.
Not applicable

1.8 What were the patient-level sociodemographic (SDS) variables that were available and analyzed in the data or sample used? For example, patient-reported data (e.g., income, education, language), proxy variables when SDS data are not collected from each patient (e.g. census tract), or patient community characteristics (e.g. percent vacant housing, crime rate).

We assessed reliability of the measures after stratifying by age, i.e., adolescent versus adult. Teen pregnancy is worthy of a separate focus because of the large potential negative impact on the life of the teen and her child(ren), and the existence of unique programs and contraceptive counseling approaches tailored to this population. To define age groups, we used the categories developed by the Center for Medicaid and CHIP Services (CMCS), i.e., individuals aged 15 through 20 years (15-20) were defined as adolescents, and individuals aged 21 through 44 years (21-44) were defined as adults.

2a2. RELIABILITY TESTING

***Note**: If accuracy/correctness (validity) of data elements was empirically tested, separate reliability testing of data elements is not required – in 2a2.1 check critical data elements; in 2a2.2 enter "see section 2b2 for validity testing of data elements"; and skip 2a2.3 and 2a2.4.*

2a2.1. What level of reliability testing was conducted? (*may be one or both levels*)

☐ **Critical data elements used in the measure** (*e.g., inter-abstractor reliability; data element reliability must address ALL critical data elements*)

☒ **Performance measure score** (e.g., *signal-to-noise analysis*)

2a2.2. For each level checked above, describe the method of reliability testing and what it tests (*describe the steps—do not just name a method; what type of error does it test; what statistical analysis was used*)

Several methods have been suggested to assess the reliability of provider-level performance measures (Adams, 2010; Scholle et al, 2008; Fung et al, 2010). These methods may focus on different facets of reliability such as consistency across time, consistency across raters or units, or variability at different levels of aggregation. The NQF has suggested a ***signal-to-noise*** approach as one way to evaluate measure reliability. According to Adams (2009), reliability can be assessed by the proportion of variance in a performance measure due to systemic differences across measured units (signal) in relation to random error (noise) within units.

When analytic units fall into a natural hierarchy (e.g. clients nested within health centers nested within health plan organizations), one can estimate multilevel variance components using hierarchical generalized linear modeling (HGLM) (Raudenbush and Bryk, 2002; Woltman et al, 2012). In this approach the within-provider regression coefficients are allowed to vary across providers as random effects. The covariance parameter for the random effect estimates the true between-provider variance after accounting for within-provider variance. HGLM methods are robust and well-developed for continuous outcomes, and have more recently been applied to binary outcomes (Ridout, 1999; Molenberghs et al, 2007).

In the present analyses, multi-level mixed models were fit to each dataset using a hierarchical SAS 9.3 GLIMMIX procedure with a log link function. Parameters were estimated by pseudo-maximum-likelihood using the Laplace method (Ene et al, 2012). Modeling proceeded in a top-down manner starting from the largest unit of aggregation; the variance component (random coefficient) was always estimated for the top level.

Reliability was then calculated as a function of the intraclass correlation (ICC) and the median number cases per unit, using the Spearman-Brown prophecy (Eijkenaar et al, 2013). ICCs are derived using the estimated variance component for the level of interest divided by the total variance (Wu et al 2012; He et al, 2014). ICCs conceptually represent the proportion of total variation accounted for by the between-provider level, and thus follows the signal-to-noise framework suggested by NQF.

The HGLM method of estimation assumes a normally distributed error component; some authors have noted that ICCs on the logit scale can be inflated under certain circumstances when population rates are near the extremes (Wu et al, 2012). To provide more conservative estimation, medians were used in the Spearman-Brown reliability formula; the use of means would tend to bias estimates upward due to one or two atypically large provider units.

Structure of the Data

Iowa Medicaid Enterprise dataset. For IME data, modeling proceeded from the level of public health region (n=6).

Louisiana Medicaid Enterprise dataset. For this dataset, modeling proceeded from the levels of public health region (n=9) and the health plan (n=5).

2a2.3. For each level of testing checked above, what were the statistical results from reliability testing? (e.*g., percent agreement and kappa for the critical data elements; distribution of reliability statistics from a signal-to-noise analysis*)

The table below shows summary results of the reliability analyses for the two state Medicaid programs at the public health region and health plan levels, stratified by three age categories (i.e., 15-20, 21-44, and 15-44) and by 3- and 60-day postpartum. The cell sizes were very small for the 3-day estimates, so we only analyzed reliability for the combined 15-44 year age group. More detailed information about the analyses at each level can be found in Tables 1-4 (appended at the end of the form).

For each level, the overall reliability was estimated using the medians as previously mentioned. ICCs, an indicator of the proportion of variance explained by the groupings, are also shown. Similar studies of hierarchical binary outcomes estimate ICCs in a typical range of .02 - .18 (Fung et al, 2010).

Iowa Medicaid Enterprise

	Level	Age group	Results		
			Median N	ICC	Reliability
3-day MOST/MOD	Region	15-44	1620	.0053	.8969
60-day MOST/MOD	Region	15-20	323	.0026	.4537
		21-44	1298	.0054	.8760
		15-44	1620	.0055	.8989
3-day LARC	Region	15-44	1620	.4301	.9992
60-day LARC	Region	15-20	323	.0025	.4445
		21-44	1298	.0039	.8357
		15-44	1620	.0044	.8772

Louisiana Medicaid

	Level	Age group	Results		
			Median N	ICC	Reliability
3-day most and moderately effective	Health plan	15-44	5096	.0022	.9190
	Region	15-44	3666	.0085	.9693
60-day most and moderately effective	Health plan	15-20	975	.0112	9172
		21-44	4151	.0066	.9650
		15-44	5096	.0072	.9736
	Region	15-20	666	.0026	.6364
		21-44	3000	.0033	.9082
		15-44	3666	.0028	.9125
3-day LARC	Health plan	15-44	5096	.0120	.9841
	Region	15-44	3666	.1677	.9986
60-day LARC	Health plan	15-20	975	.0249	.9614
		21-44	4151	.0340	.9932
		15-44	5096	.0327	.9942
	Region	15-20	666	.0381	.9634
		21-44	3000	.0271	.9882
		15-44	3666	.0299	.9912

3-day postpartum (most/mod and LARC)

The estimated reliabilities for 3-day provision of both most/moderately effective and LARC methods were high across both Medicaid programs at region and health plan levels for the 15-44 age group (the range was from .8969 to .9992). The ICCs were highly variable, ranging from 0.002 to .430. These findings were sustained when we estimated reliabilities using the minimum patient volume rather than the median (see details in tables 1-4 appended at the end of this document). It is likely that the instances of unusually high ICCs were driven by the extremely low numbers of cases and low rates of provision for LARC (less than 1% in both Medicaid programs) which may have caused instability during estimation. More typical lower ICC estimates were found for most and moderately effective methods, as might be expected given the limited variation in rates (see tables 1-4, appended at the end of this document).

60-day postpartum (most/mod and LARC)

The estimated reliabilities for 60-day provision of both most/moderately effective and LARC methods were high across both Medicaid programs at region and health plan levels for the 20-44 and 15-44 age groups (the range was from .8357 to .9942). For adolescents the results were mixed. In Louisiana at the regional level reliabilities were high among adolescents for most/moderately effective and LARC methods (.9172, .9634, respectively), and at the health plan level for LARC (.9614). However, the reliability for the adolescent age group dropped below .70 for most and moderately

effective (.4537) and LARC (.4445) methods at the regional level in Iowa; and for most and moderately effective methods (.6364) at the regional level in Louisiana. Overall, the ICCs were low in Iowa and were all below .01; In Louisiana, the ICCs were slightly higher and ranged from .003 to .038. These findings were sustained when we estimated reliabilities using the minimum patient volume rather than the median (see details in tables 1-4 appended at the end of this document).

2a2.4 What is your interpretation of the results in terms of demonstrating reliability? (i.*e., what do the results mean and what are the norms for the test conducted?*)

In performing this analysis, we attempted to provide a conservative estimate of reliability wherever possible. Using medians rather than means, and presenting the 'floor' of reliability that may be observed for the smallest units, we bracket the results with worst-case scenarios.

For the 60-day measures of postpartum contraceptive use (both most/moderately effective and LARC methods), we found that most reliabilities were close to or exceeded the .90 threshold for reporting. Our conclusion is that there is sufficient reliability for reporting contraceptive rates for the 60-day postpartum period.

We also found high levels of reliability for the 3-day measures of postpartum contraceptive use (both most/moderately effective and LARC methods). However, it is likely that these high reliabilities were influenced by the extremely low numbers of cases and low rates of provision for LARC (less than 1% in both Medicaid programs), and by the limited variation in rates of provision for most and moderately effective methods. For this reason, we conclude that there is not yet sufficient reliability to reporting on the 3-day measure. However, these measures reflect a very new clinical practice and it is not surprising that the rates of provision are so low; we are optimistic that the reliabilities will be high once the rates are higher, as has been demonstrated with the 60-day measures. Given this, we recommend that NQF endorse the 3-day measures on a 'provisional' basis.

It is commonly advised that reliability should be $\geq$ .90 for making decisions, and $\geq$ .70 for general reporting/monitoring (Eijkenaar, 2013; Adams, 2010). The Spearman-Brown prophecy allows one to test different values for ICC and patient volume per unit in order to predict expected reliability. Using an ICC value near the 20th percentile as a conservative expected value among units, we computed the minimum recommended case load for each threshold of reliability:

- For the 60 day most/moderately effective measure, we looked at the full set of ICC values for regions and plans among all age groups. Using a 20th percentile value of .0027 ICC (i.e., 80% of ICC values would be expected above this level), we would recommend that regions or plans have at least 862 patient cases for reporting rates to maintain >.70 reliability, and 3,324 cases to maintain >.90 reliability.

- For the 60-day LARC measure, we similarly grouped the set of ICC values for regions and plans among all age groups. Using a 20th percentile ICC value of .0042, we would recommend that regions or plans have at least 553 patient cases for reporting rates to maintain >.70 reliability, and 2,134 cases to maintain >.90 reliability.

- For the 3-day most/moderately effective and LARC measures, we looked at the set of ICC values for regions and plans among the 15-44 age group only (we do not recommend stratifying by teen/adult at this time). Due to the provisional nature of both measures, we used the lowest estimated ICC of .0022 for both measures as our conservative floor. We would recommend that regions or plans have at least 1,058 patient cases for reporting rates to maintain >.70 reliability, and 4,082 cases to maintain >.90 reliability.

2b2. VALIDITY TESTING

2b2.1. What level of validity testing was conducted? (*may be one or both levels*)

☐ **Critical data elements** (*data element validity must address ALL critical data elements*)

☒ **Performance measure score**

☐ **Empirical validity testing**

☒ **Systematic assessment of face validity of performance measure score as an indicator** of quality or resource use (*i.e., is an accurate reflection of performance on quality or resource use and can distinguish good from poor performance*)

2b2.2. For each level of testing checked above, describe the method of validity testing and what it tests (*describe the steps—do not just name a method; what was tested, e.g., accuracy of data elements compared to authoritative source, relationship to another measure as expected; what statistical analysis was used)*

We used a systematic process to assess the face validity of the performance measure, i.e., whether the corresponding measure scores correctly reflect the quality of care provided and adequately identify differences in quality. Nine experts with the following characteristics were identified: (1) expertise in the delivery of contraceptive services, as evidenced by employment in a clinical or managerial capacity for at least 3 years during which they delivered contraceptive services in a clinical setting (i.e., public and private family planning and primary care providers, or health administrators); and (2) expertise in the use of performance measures, as evidenced by participation in at least one effort to collect and use performance measurement data for the purpose of improving clinical services in the setting(s) in which they work. Below is the final list of experts who participated in the assessment:

1. Carol Brady, MA, Project Director, Florida Association of Healthy Start Coalitions, Inc.
2. Anne Burke, MD, Associate Professor, School of Medicine, Johns Hopkins Bayview Medical Center
   Vanessa Dalton, MD, MPH, Associate Professor, Director, Program on Women's Health Care Effectiveness Research, University of Michigan
3. Anne Dunlop, MD, MPH, Program Director, Preventive Medicine Division, Emory University School of Medicine
4. Daryn Eikner, MS, Vice President of Health Care Delivery, National Family Planning & Reproductive Health Association
5. Jan Engstrom, PhD, RN, CNM, WHNP-BC, Professor & Acting Chairperson, Department of Women, Children and Family Nursing, College of Nursing, Armour Academic Center
6. Mark Hathaway, MD, MPH, Senior Technical Advisor, Jhpiego – Johns Hopkins University
7. Michael Policar, MD, MPH, Clinical Professor of Obstetrics, Gynecology, and Reproductive Sciences, UCSF School of Medicine
8. Linda Wheal, Maternal Health Program Manager, Bureau of Quality Management, Illinois Department of Healthcare and Family Services

We contacted the selected experts to confirm consent to participate via email. Each expert panelist was sent a disclosure form to report any relevant financial or other competing interests; disclosures were compiled with brief biographies and shared with all panelists. Upon receipt of the disclosure form we sent the participant information about the measure specifications and other background information about the measure. Participants then participated in a webinar designed to provide important background information about the measure, how it is computed, the NQF endorsement process, and how the face validity assessment will be used in the application package that will be submitted to NQF. After reviewing the measure specifications and participating in the webinar the participants completed a survey (anonymous) that asked the following question about the measure:

The scores obtained from the measure as specified will provide an accurate reflection of quality and can be used to distinguish good and poor quality in contraceptive services:

1= Strongly Disagree
3=Neither Agree nor Disagree
5= Strongly Agree

ICD-10 conversion

We tested the measure specifications based on 2014 codes, but have also included the codes needed to calculate the measure using ICD-10 and 2015 NDC codes. Both sets of codes are attached. Our goal was to convert the measure to a new code set, fully consistent with the intent of the original measure. A description of how we converted from ICD-9 to ICD-10 is provided below, for each table in the measure specifications.

- **Delivery resulting in a live birth (Table PCU-A)**

These codes were identified by copying the Deliveries Value Set from NCQA's Prenatal & Postpartum Care (PPC) measure (NQF#1517), excluding extraction of products of conception retained and ectopic. In the PPC measure, these codes are used to identify live births.

- **Known miscarriage, ectopic pregnancy, stillbirth, or induced abortion (Table PCU-B)**

These codes were identified by copying the Non-live Births Value Set from NCQA's Prenatal & Postpartum Care (PPC) measure (NQF#1517), as well as non-live birth codes in "Chapter 15: Pregnancy, Childbirth and Purperium (O00-O9A)". In the PPC measure, these codes are used to identify live births.

- **Contraceptive codes (Tables PCU-C, D and E)**

We used ICD-10 online conversion tools and confirming codes in the ICD-10-CM Expert for Physicians complete official code set. They were cross-checked against a ICD-10 conversion chart for family planning services that was prepared by Dr Michael Policar, from the University of California-San Francisco, and confirmed with a clinical expert, Denise Wheeler, MS, Family Planning Director at the Iowa Department of Public Health. NDC codes for 2015 were updated by using the codes for contraception contained in the HEDIS specifications for Chlamydia screening.

2b2.3. What were the statistical results from validity testing? (*e.g., correlation; t-test*)

The mean rating for the most/moderately effective postpartum measure was 4.22 with a median of 4.5 (between Agree and Strongly Agree), range 3-5. There were 44.4% (n = 4) of respondents who strongly agreed, 33.3% (n = 3) of respondents who agreed, and 22.2% (n = 2) of respondents who neither agreed nor disagreed that the score obtained from the most/moderately effective postpartum measure, as specified, will provide an accurate reflection of quality and can be used to distinguish good and poor quality in contraceptive services.

The mean rating for the LARC postpartum measure was 3.78 with a median of 4 (Agree), range 1-5. There were 33.3% (n = 3) of respondents who strongly agreed, 33.3% (n = 3) of respondents who agreed, 22.2% (n = 2) of respondents who neither agreed nor disagreed, and 11.1% (n = 1) of respondents who strongly disagreed that the scores obtained from the LARC postpartum measure, as specified, will provide an accurate reflection of quality and can be used to distinguish good and poor quality in contraceptive services.

One respondent replied that he or she "feels STRONGLY that the adoption of these measures will promote providers' and practices' attention to reproductive planning and contraceptive care as part and parcel of women's primary health care." One respondent replied that he or she thinks that "the proposed measures are valid measures of quality contraceptive care for healthy women" and "the proposed measures are valid indicators of contraceptive care in postpartum women who are NOT breastfeeding." However, he or she thought, "the measures are of questionable validity in postpartum women who are breastfeeding. Breastfeeding women are being advised not to accept hormonal contraceptives during the first months after birth while exclusively breastfeeding. Postpartum women are specifically advised against the early administration of normal contraceptives." He or she thought "that although providers may offer highly effective contraceptives to postpartum women and be proficient within the administration and management of the contraceptive, more women will refuse the contraceptives. Also of concern is that some clinicians do not want to be perceived as undermining breastfeeding, so they may not broach the subject out of fear of being

labeled 'anti-breastfeeding.' Additionally, the Academy of Breastfeeding Medicine's statement on contraceptive choice during breastfeeding makes providers reluctant to offer or administer the most effective hormonal methods, especially in the early postpartum period. " He or she noted, "This is truly a situation in which two major public health initiatives collide." One respondent thought the most/moderately effective postpartum measure should extend to 90 days postpartum, not 60 days.

2b2.4. What is your interpretation of the results in terms of demonstrating validity? (i.*e., what do the results mean and what are the norms for the test conducted?*)

We think that the responses to the face validity assessment indicate that the measure assessing provision of most/moderately effective methods will provide an accurate reflection of quality and can be used to distinguish good and poor quality in contraceptive services.

There was less support for the measure assessing provision of LARC in the postpartum period. There were also concerns expressed about the validity of the measure for breastfeeding women, especially within 90 days of delivery. Despite these reservations, OPA suggests that the LARC measure be approved on a provisional basis so that more experience can be gained using the measures. We will convene an expert panel within 2 years of endorsement to review results and make suggestions for revising the measure, as needed.

_____

2b3. EXCLUSIONS ANALYSIS

NA ☐ no exclusions — *skip to section 2b4*

2b3.1. Describe the method of testing exclusions and what it tests (*describe the steps—do not just name a method; what was tested, e.g., whether exclusions affect overall performance scores; what statistical analysis was used*)

Exclusions were not formally tested. The rationale for exclusion was due to the fact that some women are not at risk of unintended pregnancy due to infecundity or pregnancy.

2b3.2. What were the statistical results from testing exclusions? (*include overall number and percentage of individuals excluded, frequency distribution of exclusions across measured entities, and impact on performance measure scores*)

The table below shows the number of women excluded in each of the two datasets, presented by the reason for exclusion.

		Number of women	
		IME, 2013	Louisiana Medicaid, 2014
All live births to women 15-44 years of age in the measurement year		14,776	42,603
Exclusions	Deliveries that did not end in a live birth (miscarriage, ectopic, stillbirth or induced abortion)	240	1,481
	Live births that occurred in the last 2 months of the measurement year	2,167	7,142
Number of live births to women 15-44 years of age, after exclusions		12,369*	33,980

* 844 women were not included in the final analysis for Iowa because of missing data about the health plan to which they belonged. This resulted in a sample of 11,525 women.

2b3.3. What is your interpretation of the results in terms of demonstrating that exclusions are needed to prevent unfair distortion of performance results? (*i.e., the value outweighs the burden of increased data collection and analysis.*

*Note: **If patient preference is an exclusion**, the measure must be specified so that the effect on the performance score is transparent, e.g., scores with and without exclusion*)

When combined, the total number of exclusions in each of the two data sets comprised 16% (IME) and 20% (Louisiana Medicaid) of all postpartum women 15-44 years of age. The number of women excluded will have a noticeable impact on the rates, and excluding these categories of women will be important to reassure providers that the measure is as 'fair' in terms of only including women that providers will have had an opportunity to care for. For these reasons, we believe that the burden of applying the exclusion criteria is outweighed by the benefits of doing so.

2b4. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES

If not an intermediate or health outcome, or PRO-PM, or resource use measure, skip to section 2b5.

2b4.1. What method of controlling for differences in case mix is used?

☒ **No risk adjustment or stratification**

☐ **Statistical risk model with** Click here to enter number of factors **risk factors**

☐ **Stratification by** Click here to enter number of categories **risk categories**

☐ **Other,** Click here to enter description

2b4.2. If an outcome or resource use measure is not risk adjusted or stratified, provide rationale and analyses to demonstrate that controlling for differences in patient characteristics (case mix) is not needed to achieve fair comparisons across measured entities.

We do not believe that risk adjustment is justified. Although there are [possible] variations in contraceptive use by socio-demographic characteristics, the reason for those patterns is based on modifiable clinical and programmatic considerations rather than differing biological responses to contraception. Although providers may see some local variations by socio-demographic characteristics, we do not believe that these differences will be maintained if contraceptive services are offered in a client-centered manner, as defined by CDC-OPA recommendations for providing quality family planning services (CDC-OPA, 2014).

A special (unpublished) analysis of data from the Pregnancy Risk Assessment Monitoring System (PRAMS), 2011-2012, was conducted to explore disparities in the use of most and moderately effective and LARC methods of contraception (see table below). This analysis suggests that there are statistically significant differences by age, marital status and some income categories for use of most and moderately effective methods and LARC methods. However, there were no significant differences by race/ethnicity, and most income categories.

Percentage of women 15-44 years of age with a recent live birth that used a most/moderately effective or a LARC method of contraception, Pregnancy Risk Assessment Monitoring System (PRAMS)* 2011-2012

	Sample size (unweighted)	Most or moderately effective methods Percent [95% Confidence Limits]	LARC methods Percent [95% Confidence Limits]
Overall	146,648	49.9 [49.1,50.6]	13.6
Age			
<19	13,141	60.1[57.4,62.7]	23.1 [21.0,25.4]
20-29	74,925	52.9 [51.9,53.9]	15.8 [15.1,16.5]
>30	58,575	44.6 [43.5,45.6]	9.5 [8.9,10.1]
Race/ethnicity			
NH White	73,045	51.2 [50.3,52.1]	14.0 [13.3,14.6]

NH Black	22,389	53.9 [51.8,56.1]	13.1 [11.7,14.7]
Hispanic	23,558	51.4 [49.7,53.0]	15.0 [13.9,16.1]
Marital status			
Married	87,515	56.6 [55.4,57.8]	16.4 [15.6,17.3]
Not married	59,026	45.9 [45.0,46.7]	11.9 [11.4,12.5]
Federal poverty level			
<100	39,043	42.8 [40.8,44.8]	12.2 [11.0,13.4]
100-199	20,160	41.2 [38.5,43.8]	13.5 [11.9,15.3]
200-399	18,681	36.8 [34.2,39.4]	9.3 [8.0,10.7]
400-499	8,322	38.6 [34.4,42.9]	9.4 [7.4,12.0]
500+	60,442	52.1 [51.3,53.0]	14.1 [13.5,14.7]

* The following PRAMS reporting sites were included in the analysis: AK, AR, CO, DE, GA, HI, IL, MA, MD, ME, MI, MN, MO, NE, NJ, NM, NY, OH, OK, OR, PA, RI, TN, UT, VT, WA, WI, WV, WY, NYC.

2b4.3. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors (clinical factors or sociodemographic factors) used in the statistical risk model or for stratification by risk (*e.g., potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of p<0.10; correlation of x or higher; patient factors should be present at the start of care*)

We recommended stratifying the client population by age so that rates for adolescents can be tracked separately from those for adult women. We propose this stratification for purposes of QI but not as a method of risk-adjustment. Teen pregnancy is worthy of a separate focus because of the large potential negative impact on the life of the teen and her child(ren), and the existence of unique programs and contraceptive counseling approaches tailored to this population.

2b4.4a. What were the statistical results of the analyses used to select risk factors?

Not applicable.

2b4.4b. Describe the analyses and interpretation resulting in the decision to select SDS factors (e.g. prevalence of the factor across measured entities, empirical association with the outcome, contribution of unique variation in the outcome, assessment of between-unit effects and within-unit effects)

Not applicable.

2b4.5. Describe the method of testing/analysis used to develop and validate the adequacy of the statistical model or stratification approach (*describe the steps—do not just name a method; what statistical analysis was used*)

Provide the statistical results from testing the approach to controlling for differences in patient characteristics (case mix) below.
If stratified, skip to 2b4.9

2b4.6. Statistical Risk Model Discrimination Statistics (*e.g., c-statistic, R-squared*)**:**

Not applicable.

2b4.7. Statistical Risk Model Calibration Statistics (*e.g., Hosmer-Lemeshow statistic*):

Not applicable.

2b4.8. Statistical Risk Model Calibration – Risk decile plots or calibration curves:

Not applicable.

2b4.9. Results of Risk Stratification Analysis:

Not applicable.

2b4.10. What is your interpretation of the results in terms of demonstrating adequacy of controlling for differences in patient characteristics (case mix)? (i*.e., what do the results mean and what are the norms for the test conducted*)
Not applicable.

2b4.11. Optional Additional Testing for Risk Adjustment (*not required, but would provide additional support of adequacy of risk model, e.g., testing of risk model in another data set; sensitivity analysis for missing data; other methods that were assessed*)

Not applicable.

_____

2b5. IDENTIFICATION OF STATISTICALLY SIGNIFICANT & MEANINGFUL DIFFERENCES IN PERFORMANCE

2b5.1. Describe the method for determining if statistically significant and clinically/practically meaningful differences in performance measure scores among the measured entities can be identified (*describe the steps—do not just name a method; what statistical analysis was used? Do not just repeat the information provided related to performance gap in 1b)*

Due to the fact that our dataset represents a census, all rates are assumed to reflect 'true' rates by unit for the data year. Non-sampling error (such as coding or measurement error) is not estimable given our limited access to the claims data and processes. Thus we do not present any confidence intervals for inferential testing results. These assumed-true differences in rates must therefore be evaluated based on practical or clinically meaningful impact.

2b5.2. What were the statistical results from testing the ability to identify statistically significant and/or clinically/practically meaningful differences in performance measure scores across measured entities? (e.g., *number and percentage of entities with scores that were statistically significantly different from mean or some benchmark, different from expected; how was meaningful difference defined*)

The table below summarizes rates at each level. As noted above, since our data contain the entirety of the defined population, estimation of sampling error and related inferential statistics such as confidence intervals are not applicable. More detailed information about the variation of rates by unit within each level can be found in Tables 1-4, which are appended at the end of this document.

			Most and moderately effective methods Mean (range)	
	Level	Age	Iowa Medicaid, 2013	Louisiana Medicaid, 2014
Within 3 days of delivery	Health plan	15-44	N.A.	.096 (.082-.127)
	Public health region	15-44	.106 (.085-.123)	.097 (.083-.105)
Within 60 days of delivery	Health plan	15-20	N.A.	.416 (.362-.451)
		21-44		.438 (.385-.470)
		15-44		.434 (.382-.462)

	Public health region	15-20	.416 (.354-.468)	.427 (.348-.487)
		21-44	.426 (.388-.480)	.453 (.388-.490)
		15-44	.424 (.382-.474)	.448 (.381-482)

			LARC methods Mean (range)	
	Level	Age	Iowa Medicaid, 2013	Louisiana Medicaid, 2014
Within 3 days of delivery	Health plan	15-44	N.A.	.003 (.001-.007)
	Public health region	15-44	.004 (0 - .010)	.003 (001-.004)
Within 60 days of delivery	Health Plan	15-20	N.A.	.114 (.050-.155)
		21-44		.088 (.041-.119)
		15-44		.093 (.043-.125)
	Public health region	15-20	.144 (.102-.169)	.117 (.062-.133)
		21-44	.099 (.084-.121)	.092 (.045-.112)
		15-44	.107 (.087-.124)	.097 (.048-.115)

2b5.3. What is your interpretation of the results in terms of demonstrating the ability to identify statistically significant and/or clinically/practically meaningful differences in performance across measured entities? (i*.e., what do the results mean in terms of statistical and meaningful differences?*)

There are large differences in rates across both levels for the 60-day measures of postpartum contraception. For example, the rates for most and moderately effective methods ranged by region in Iowa (from 35% to 48%) and Louisiana (from 35% to 49%), and across health plans in Louisiana (from 36% to 47%). There were also substantial differences in provision of LARC methods across regions in Iowa (from 8% to 17%) and Louisiana (from4% to 13%), and across health plans in Louisiana (from 4% to 16%). These differences suggest that it will be possible to identify meaningful differences in performance across measured entities for the 60-day measures.

There was less variation for the 3-day postpartum measures. For example, the rates for provision of most and moderately effective methods ranged by region in Iowa (9% to 12%) and Louisiana (from 8 to 11%), and across health plans in Louisiana (from 8 to 13%). There was than one percentage point difference in provision of LARC methods across all reporting units. The provision of contraception in the immediate postpartum period is a relatively new clinical practice and payment strategies are only now being implemented, so these low rates are not surprising. However, we expect to see notable increases in the rates over the coming years.

2b6. COMPARABILITY OF PERFORMANCE SCORES WHEN MORE THAN ONE SET OF SPECIFICATIONS
***If only one set of specifications, this section can be skipped*.**

Not applicable.

Note: *This item is directed to measures that are risk-adjusted (with or without SDS factors)* ***OR*** *to measures with more than one set of specifications/instructions (e.g., one set of specifications for how to identify and compute the measure from medical record abstraction and a different set of specifications for claims or eMeasures). It does not apply to measures that use more than one source of data in one set of specifications/instructions (e.g., claims data to identify the denominator and medical record abstraction for the numerator).* ***Comparability is not required when comparing performance scores with and without SDS factors in the risk adjustment model. However, if comparability is not demonstrated for measures with more than one set of specifications/instructions, the different specifications (e.g., for medical records vs. claims) should be submitted as separate measures.***

2b6.1. Describe the method of testing conducted to compare performance scores for the same entities across the different data sources/specifications (*describe the steps—do not just name a method; what statistical analysis was used*)

Not applicable.

2b6.2. What were the statistical results from testing comparability of performance scores for the same entities when using different data sources/specifications? (*e.g., correlation, rank order*)

Not applicable.

2b6.3. What is your interpretation of the results in terms of the differences in performance measure scores for the same entities across the different data sources/specifications? (i.*e., what do the results mean and what are the norms for the test conducted*)

Not applicable.

2b7. MISSING DATA ANALYSIS AND MINIMIZING BIAS

2b7.1. Describe the method of testing conducted to identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias (*describe the steps—do not just name a method; what statistical analysis was used*)

The data source for this measure is claims data. Due to the nature of claims data (i.e., for billing purposes), there is typically very little missing data; further, it is difficult to ascertain when claims data is or is not missing.

2b7.2. What is the overall frequency of missing data, the distribution of missing data across providers, and the results from testing related to missing data? (*e.g., results of sensitivity analysis of the effect of various rules for missing data/nonresponse; if no empirical sensitivity analysis, identify the approaches for handling missing data that were considered and pros and cons of each*)

Not addressed due to the nature of claims data.

2b7.3. What is your interpretation of the results in terms of demonstrating that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias**?** (i.*e., what do the results mean in terms of supporting the selected approach for missing data and what are the norms for the test conducted; if no empirical analysis, provide rationale for the selected approach for missing data*)

Not applicable.

References

Adams JL, Mehrotra A, Thomas JW, McGlynn EA (2010). Physician cost profiling-reliability and risk of misclassification. New England Journal of Medicine; 362:1014–1021.

Adams JL (2009). The Reliability of Provider Profiling: A Tutorial. Rand Health, Santa Monica, CA. Last accessed online at: http://www.rand.org/pubs/technical_reports/TR653.

Eijkenaar F, van Vliet RC. (2013). Profiling individual physicians using administrative data from a single insurer: variance components, reliability, and implications for performance improvement efforts. Med Care. 51(8):731-9

Ene, M, Leighton, EA, Blue, GL, Bell, BA (2015). Models for categorical data using SAS PROC GLIMMIX. Sas Global Forum, Dallas TX, April 26-29, Paper 3430. Last accessed online at: http://support.sas.com/resources/papers/proceedings15/3430-2015.pdf.

Fung, V, Schmittdiel, JA, Fireman, B, Meer, A, Thomas, S, Hsu, J, Selby, JV (2010). Meaningful variation in performance: A systematic literature review. Medical Care, 48(2): 140-148.

He, Y, Selck, F, and Normand, SL (2014). On the accuracy of classifying hospitals on their performance measures. Statistics in Medicine; 33: 1081-1103.

Martin JA, Hamilton BE, Osterman MJ, Curtin SC, Matthews TJ. (2015) Births: final data for 2013. Natl Vital Stat Rep; 64(1):1-65.

Molenberghs G, Laenen A, Vengeneugden T (2007). Estimating reliability and generalizability from hierarchical biomedical data. Journal of Biopharmaceutical Statistics, 17(4): 595-627.

National Committee for Quality Assurance (NCQA). HEDIS 2015: Healthcare Effectiveness Data and Information Set. Vol. 1, narrative. Washington (DC): National Committee for Quality Assurance (NCQA); Retrieved from: http://www.qualitymeasures.ahrq.gov/content.aspx?id=48621

NQF 2009. National Quality Forum, Health Information Technology Expert Panel II- Health IT Enablement of Quality Measurement. Washington DC.

Raudenbush SW and Bryk AS. *Hierarchical Linear Models: Applications and Data Analysis Methods (2nd Edition)*. Sage Publications, 2002: Thousand Oaks.

Ridout MS, Demetrio CGB, Firth D (1999). Estimating intraclass correlation for binary data. Biometrics, 55(1): 137-148.

Romero L, Pazol K, Warner L, Gavin L, Moskosky S, Besera G, et al. (2015). Vital Signs: Trends in Use of Long-Acting Reversible Contraception Among Teens Aged 15–19 Years Seeking Contraceptive Services — United States, 2005–2013. MMWR Weekly Report April 10, 2015 / 64(13);363-369

Scholle SH, Roski J, Adams JL, Dunn DL, Kerr EA, Pilleterre-Dugan D, Jensen RE (2008). Benchmarking physician performance: reliability of individual and composite measures. Am J of Managed Care, 14(12): 833-838.

Sedgh G, Finer LB, Bankole A, Eilers MA, Singh S. (2015) Adolescent pregnancy, birth, and abortion rates across countries: levels and recent trends. J Adolesc Health; 56(2):223-30.

Woltman H, Feldstain A, MacKay JC, Rocchi M. (2012). An introduction to hierarchical linear modeling. Tutorials in Quantitative Methods for Psychology; 8(1):52–69.

Wu S, Crespi CM, Wong WK (2012). Comparison of methods for estimating the intraclass correlation coefficient for binary responses in cancer prevention cluster randomized trials. Contemp Clin Trials, 33(5): 869-880.

Table 1. Rates and reliabilities for use of most and moderately effective contraceptive methods in the postpartum period, Iowa Medicaid Enterprise, 2013, by public health region

3 DAY POSTPARTUM – MOST/MOD

Public Health Region	MEM_MethodMost: 15 to <21 Years	MEM_MethodMost: 21 to 45 years	MEM_MethodMost: all age groups					
			Not Used	Used Most/Mod	Total N	Rate		
1			3103	290	3393	0.085		
2			823	113	936	0.121		
3			1546	180	1726	0.104		
4	N.A.	N.A.	1099	124	1223	0.101		
5			1328	186	1514	0.123		
6			2402	331	2733	0.121		
Total or Mean			**10301**	**1224**	**11525**	**0.106**		
						VarL1	**ICC**	Region Reliability (Var L1)
Median Patient Volume				Median n	1620	0.01767	0.0053	**0.8969**
Minimum Patient Volume (Floor)				Min n	936	0.01767	0.0053	0.8341

60 DAY POSTPARTUM - MOST/MOD

Public Health Region	MEM_MethodMost: 15 to <21 Years						MEM_MethodMost: 21 to 45 years						MEM_MethodMost: all age groups, 15-44 years					
	Not Used	Used Most/ Mod	Total N	Rate			Not Used	Used Most/ Mod	Total N	Rate			Not Used	Used Most/ Mod	Total N	Rate		
1	363	243	606	0.401			1689	1098	2787	0.394			2052	1341	3393	0.395		
2	81	64	145	0.441			411	380	791	0.480			492	444	936	0.474		
3	203	111	314	0.354			864	548	1412	0.388			1067	659	1726	0.382		
4	138	106	244	0.434			526	453	979	0.463			664	559	1223	0.457		
5	176	155	331	0.468			632	551	1183	0.466			808	706	1514	0.466		
6	290	213	503	0.423			1265	965	2230	0.433			1555	1178	2733	0.431		
Total or Mean	**1251**	**892**	**2143**	**0.416**			**5387**	**3995**	**9382**	**0.426**			**6638**	**4887**	**11525**	**0.424**		
				VarL1	**ICC**	Region Reliability (Var L1)				**VarL1**	**ICC**	Region Reliability (Var L1)				**VarL1**	**ICC**	Region Reliability (Var L1)
Median Patient Volume		Median n	322.5	0.00847	0.0026	**0.4537**		Median n	1298	0.01079	0.0054	**0.8760**		Median n	1620	0.01081	0.0055	**0.8989**
Minimum Patient Volume (Floor)		Min n	145	0.00847	0.0026	0.2719		Min n	791	0.01079	0.0054	0.8116		Min n	936	0.01081	0.0055	0.8370

Table 2. Rates and reliabilities for use of LARC methods in the postpartum period, Iowa Medicaid Enterprise, 2013, by public health region

3 DAY POSTPARTUM - LARC

Public Health Region	LARCMeasure: 15 to <21 Years		LARCMeasure: 21 to 45 years		LARCMeasure: all age groups					
					Not Used	Used LARC	Total N	Rate		
1					3392	1	3393	0.000		
2					935	1	936	0.001		
3					1725	1	1726	0.001		
4	N.A.		N.A.		1223	0	1223	0.000		
5					1502	12	1514	0.008		
6					2705	28	2733	0.010		
Total or Mean					**11482**	**43**	**11525**	**0.004**		
								VarL1	**ICC**	Region Reliability (Var L1)
Median Patient Volume						Median n	1620	2.483	0.4301	**0.9992**
Minimum Patient Volume (Floor)						Min n	936	2.483	0.4301	0.9986

60 DAY POSTPARTUM - LARC

Public Health Region	LARC Measure: 15-20 Years						LARC Measure: 21-44 years						LARC Measure: all age groups, 15-44 years					
	Not Used	Used LARC	Total N	Rate			Not Used	Used LARC	Total N	Rate			Not Used	Used LARC	Total N	Rate		
1	519	87	606	0.144			2534	253	2787	0.091			3053	340	3393	0.100		
2	125	20	145	0.138			695	96	791	0.121			820	116	936	0.124		
3	282	32	314	0.102			1293	119	1412	0.084			1575	151	1726	0.087		
4	211	33	244	0.135			894	85	979	0.087			1105	118	1223	0.096		
5	279	52	331	0.157			1058	125	1183	0.106			1337	177	1514	0.117		
6	418	85	503	0.169			1980	250	2230	0.112			2398	335	2733	0.123		
Total or Mean	**1834**	**309**	**2143**	**0.144**			**8454**	**928**	**9382**	**0.099**			**10288**	**1237**	**11525**	**0.107**		
				VarL1	**ICC**	Region Reliability (Var L1)				**VarL1**	**ICC**	Region Reliability (Var L1)				**VarL1**	**ICC**	Region Reliability (Var L1)
Median Patient Volume		Median n	325	0.00816	0.0025	**0.4445**		Median n	1298	0.0129	0.0039	**0.8357**		Median n	1620	0.01451	0.0044	**0.8772**
Minimum Patient Volume (Floor)		Min n	145	0.00816	0.0025	0.2646		Min n	791	0.0129	0.0039	0.7562		Min n	936	0.01451	0.0044	0.8050

Table 3. Rates and reliabilities for use of most and moderately effective contraceptive methods in the postpartum period, Louisiana Medicaid, 2014, by region and health plan

3-DAY POSTPARTUM / MOST/MOD

Public Health Region	MEM_MethodMost: 15 to <21 Years	MEM_MethodMost: 21 to 45 years	MEM_MethodMost: both age groups, 15-44					
			Not Used	Used Most-Mod	Total N	Rate		
1	N.A.	N.A.	6123	537	6660	0.081		
2			4265	375	4640	0.081		
3			2698	337	3035	0.111		
4			4076	524	4600	0.114		
5			2041	209	2250	0.093		
6			1988	290	2278	0.127		
7			3505	417	3922	0.106		
8			2662	267	2929	0.091		
9			3366	300	3666	0.082		
Total or Mean			**30724**	**3256**	**33980**	**0.096**		
						VarL1	**ICC**	Region Reliability (Var L1)
Median Patient Volume				Median n	3666	0.02831	0.0085	**0.9693**
Minimum Patient Volume (Floor)				Min n	2250	0.02831	0.0085	0.9509

Health Plan		MEM_MethodMost: 15 to <21 Years	MEM_MethodMost: 21 to 45 years	MEM_MethodMost: both age groups, 15-44					
				Not Used	Used Most/Mod	Total N	Rate		
MCO1	Prepaid	N.A.	N.A.	4625	471	5096	0.092		
MCO2	Prepaid			4555	410	4965	0.083		
MCO3	Prepaid			4367	448	4815	0.093		
MCO4	Shared			6309	731	7040	0.104		
MCO5	Shared			8940	1045	9985	0.105		
Total or Mean				**28796**	**3105**	**31901**	**0.097**		
							VarL2	**ICC**	Health plan Reliability (Var L2)
Reliability Based on Median Patient Volume					Median n	5096	0.007329	0.0022	**0.9190**
Calculated Based on Minimum					Min n	4815	0.007329	0.0022	0.9147

Patient Volume (Floor)						

60-DAY POSTPARTUM / MOST/MOD

Public Health Region	MEM_MethodMost: 15 to <21 Years						MEM_MethodMost: 21 to 45 years						MEM_MethodMost: both age groups, 15-44					
	Not Used	Used Most/ Mod	Total N	Rate			Not Used	Used Most/ Mod	Total N	Rate			Not Used	Used Most/ Mod	Total N	Rate		
1	591	335	926	0.362			3528	2206	5734	0.385			4119	2541	6660	0.382		
2	421	336	760	0.442			2146	1734	3880	0.447			2570	2070	4640	0.446		
3	318	222	540	0.411			1322	1173	2495	0.470			1640	1395	3035	0.460		
4	517	424	941	0.451			1957	1702	3659	0.465			2474	2126	4600	0.462		
5	284	205	485	0.423			1010	763	1765	0.432			1294	968	2250	0.430		
6	305	193	501	0.385			981	789	1770	0.445			1286	982	2268	0.433		
7	456	304	760	0.400			1701	1461	3162	0.462			2157	1765	3922	0.450		
8	322	246	568	0.433			1223	1100	2323	0.473			1546	1346	2892	0.465		
9	376	290	666	0.435			1417	1260	2677	0.470			1793	1550	3343	0.463		
Total or Mean	**3592**	**2555**	**6147**	**0.416**			**15645**	**12188**	**27833**	**0.438**			**19237**	**14743**	**33980**	**0.434**		
					VarL1	ICC	Region Reliability (Var L1)				VarL1	ICC	Region Reliability (Var L1)				VarL1	ICC	Region Reliability (Var L1)
Median Patient Volume		Median n	666	0.008648	0.026	**0.6364**		Median n	3000	0.01085	0.033	**0.9082**		Median n	3666	0.009354	0.028	**0.9125**
Minimum Patient Volume (Floor)		Min n	485	0.008648	0.026	0.5604		Min n	1765	0.01085	0.033	0.8534		Min n	2250	0.009354	0.028	0.8648

Health Plan	MEM_MethodMost: 15 to <21 Years						MEM_MethodMost: 21 to 45 years						MEM_MethodMost: both age groups, 15-44					
	Not Used	Used Most/ Mod	Total N	Rate			Not Used	Used Most/ Mod	Total N	Rate			Not Used	Used Most/ Mod	Total N	Rate		
MCO1 Prepaid	5409	405	945	0.429			23997	1752	41512	0.422			29396	2157	5096	0.423		
MCO2 Prepaid	614	327	941	0.348			2461	1563	4024	0.388			3075	1890	4965	0.381		
MCO3 Prepaid	602	373	975	0.383			2166	1674	3840	0.436			2768	2047	4815	0.425		
MCO4 Shared	662	629	1291	0.487			3004	2745	5749	0.477			3666	3374	7040	0.479		
MCO5 Shared	947	769	1716	0.448			4242	4048	8290	0.488			5189	4817	9985	0.482		
Total or Mean	**3365**	**2503**	**5868**	**0.427**			**14251**	**11782**	**26033**	**0.453**			**17616**	**14285**	**31901**	**0.448**		
					VarL2	ICC	Health plan Reliability (Var L2)				VarL2	ICC	Health plan Reliability (Var L2)				VarL2	ICC	Health plan Reliability (Var L2)
Reliability Based on Median Patient Volume		Median n	975	0.03074	0.0112	**0.9172**		Median n	4151	0.02186	0.0066	**0.9650**		Median n	5096	0.02385	0.0072	**0.9736**

Calculated Based on Minimum Patient Volume (Floor)	Min n	941	0.0374	0.0112	0.9145	Min n	3840	0.02186	0.0066	0.9623	Min n	4815	0.02385	0.0072	0.9721

Table 4. Rates and reliabilities for use of LARC methods in the postpartum period, Louisiana Medicaid, 2014, by region and health plan

3-DAY POSTPARTUM LARC

Public Health Region	LARCMeasure: 15 to <21 Years	LARCMeasure: 21 to 45 years	LARCMeasure: both age groups, 15-44					
			Not Used	Used LARC	Total N	Rate		
1			6644	16	6660	0.002		
2			4628	12	4640	0.003		
3			3033	2	3035	0.001		
4			4595	5	4600	0.001		
5	N.A.	N.A.	2237	13	2250	0.006		
6			2277	1	2278	0.000		
7			3894	28	3922	0.007		
8			2921	8	2929	0.003		
9			3664	2	3666	0.001		
Total or Mean			**33893**	**87**	**33980**	**0.003**		
						VarL1	**ICC**	Region Reliability (Var L1)
Median Patient Volume				Median n	3666	0.663	0.1677	**0.9986**
Minimum Patient Volume (Floor)				Min n	2250	0.663	0.1677	0.9978

Health Plan		LARCMeasure: 15 to <21 Years	LARCMeasure: 21 to 45 years	LARCMeasure: both age groups, 15-44					
				Not Used	Used LARC	Total N	Rate		
MCO1	Prepaid			5078	18	5096	0.004		
MCO2	Prepaid			4958	7	4965	0.001		
MCO3	Prepaid			4801	14	4815	0.003		
MCO4	Shared			7014	26	7040	0.004		
MCO5	Shared	N.A	N.A	9964	21	9985	0.002		
Total or Mean				**31815**	**86**	**31901**	**0.003**		
							VarL2	**ICC**	Health plan Reliability (Var L2)
Reliability Based on Median Patient Volume					Median n	5096	0.03997	0.0120	**0.9841**
Calculated Based on Minimum Patient Volume (Floor)					Min n	4815	0.03997	0.0120	0.9832

60-DAY POSTPARTUM LARC

Public Health Region	LARCMeasure: 15 to <21 Years						LARCMeasure: 21 to 45 years						LARCMeasure: both age groups, 15-44					
	Not Used	Used LARC	Total N	Rate			Not Used	Used LARC	Total N	Rate			Not Used	Used LARC	Total N	Rate		
1	810	116	926	0.125			5253	481	5734	0.084			6063	597	6600	0.090		
2	249	118	540	0.085			1922	461	2480	0.09			2727	579	3040	0.025		
3	387	47	940	0.067			3434	225	3659	0.061			4302	272	4635	0.090		
4	742	64	1048	0.115			2515	234	2749	0.085			3200	298	3500	0.065		
5	947	56	1005	0.055			1797	168	1965	0.085			2621	224	2850	0.000		
6	664	25	1076	0.140			2805	72	3177	0.041			3481	97	3878	0.043		
7	749	113	1560	0.097			2121	341	2462	0.108			2634	454	2922	0.160		
8	058	78	866	0.127			4327	218	6102	0.032			3326	296	2901	0.010		
9	5	81	666	0.122			4125	259	6000	0.066			2633	340	6693	0.000		
Total or Mean	**5449**	**698**	**6147**	**0.114**			**25374**	**2459**	**27833**	**0.088**			**30823**	**3157**	**33980**	**0.093**		
					Var L1	**ICC**	**Region Reliability (Var L1)**				**Var L1**	**ICC**	**Region Reliability (Var L1)**				**Var L1**	**ICC**	**Region Reliability (Var L1)**
Median Patient Volume		Median n	666	0.13002	0.0381	**0.9634**		Median n	3000	0.09167	0.0271	**0.9882**		Median n	3666	0.1013	0.0299	**0.9912**
Minimum Patient Volume (Floor)		Min n	485	0.13002	0.0381	0.9505		Min n	1765	0.09167	0.0271	0.9801		Min n	2250	0.1013	0.0299	0.9858

Health Plan	LARCMeasure: 15 to <21 Years						LARCMeasure: 21 to 45 years						LARCMeasure: both age groups, 15-44					
	Not Used	Used LARC	Total N	Rate			Not Used	Used LARC	Total N	Rate			Not Used	Used LARC	Total N	Rate		
MCO1 Prepaid	831	114	945	0.121			3795	356	4151	0.086			4626	470	5096	0.092		
MCO2 Prepaid	883	58	941	0.062			3843	181	4024	0.045			4726	239	4965	0.048		
MCO3 Prepaid	860	115	975	0.118			3492	348	3840	0.091			4352	463	4815	0.096		
MCO4 Shared	011	179	125	0.139			5151	598	5749	0.104			6263	777	7040	0.110		
MCO5 Shared	1495	221	1716	0.129			4573	924	6902	0.132			4085	1145	8501	0.150		
Total or Mean	**5181**	**687**	**5868**	**0.117**			**23626**	**2407**	**26033**	**0.092**			**28807**	**3094**	**31901**	**0.097**		
					Var L2	**ICC**	**Health plan Reliability (Var L2)**				**Var L2**	**ICC**	**Health plan Reliability (Var L2)**				**Var L2**	**ICC**	**Health plan Reliability (Var L2)**
Reliability Based on Median Patient		Median n	975	0.08402	0.0249	**0.9614**		Median n	4151	0.11157	0.0340	**0.9932**		Median n	5096	0.1113	0.0327	**0.9942**

Volume															
Calculated Based on Minimum Patient Volume (Floor)	Min n	941	0.08402	0.0249	0.9600	Min n	3840	0.1157	0.0340	0.9926	Min n	4815	0.1113	0.0327	0.9939

3. Feasibility

Extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

3a. Byproduct of Care Processes

For clinical measures, the required data elements are routinely generated and used during care delivery (e.g., blood pressure, lab test, diagnosis, medication order).

3a.1. **Data Elements Generated as Byproduct of Care Processes.**

Coded by someone other than person obtaining original information (e.g., DRG, ICD-9 codes on claims)

If other:

3b. Electronic Sources

The required data elements are available in electronic health records or other electronic sources. If the required data are not in electronic health records or existing electronic sources, a credible, near-term path to electronic collection is specified.

3b.1. **To what extent are the specified data elements available electronically in defined fields?** *(i.e., data elements that are needed to compute the performance measure score are in defined, computer-readable fields)*

ALL data elements are in defined fields in electronic claims

3b.2. **If ALL the data elements needed to compute the performance measure score are not from electronic sources, specify a credible, near-term path to electronic capture, OR provide a rationale for using other than electronic sources.**

3b.3. **If this is an eMeasure, provide a summary of the feasibility assessment in an attached file or make available at a measure-specific URL.**

Attachment:

3c. Data Collection Strategy

Demonstration that the data collection strategy (e.g., source, timing, frequency, sampling, patient confidentiality, costs associated with fees/licensing of proprietary measures) can be implemented (e.g., already in operational use, or testing demonstrates that it is ready to put into operational use). For eMeasures, a feasibility assessment addresses the data elements and measure logic and demonstrates the eMeasure can be implemented or feasibility concerns can be adequately addressed.

3c.1. **Describe what you have learned/modified as a result of testing and/or operational use of the measure regarding data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, time and cost of data collection, other feasibility/implementation issues.**

IF a PRO-PM, consider implications for both individuals providing PROM data (patients, service recipients, respondents) and those whose performance is being measured.

The measure has been piloted in Iowa and Louisiana Medicaid programs, and may be reported by up to 13 state Medicaid programs that are funded by the Center for Medicaid and CHIP Services (CMCS) to report from 2015-2018 on the measure as part of the Maternal and Infant Health initiative (see below). Overall, these experiences have confirmed that the measures can be feasibly and meaningfully calculated using existing claims data. Some other lessons learned are summarized below:

- The use of claims data results in a sample that is nearly identical to use of the birth record. There were 14,757 live births identified in the Medicaid claims data using the SAS code written for the measure specifications and 15,212 Medicaid live births were linked to the birth certificate in 2013, which is only a difference of 455 live births (i.e., 3%).
- State-specific NDC codes may need to be added: Several of the CMCS-funded states found that some NDC codes were missing from the measure specifications, especially for oral contraception. Obtaining a comprehensive list of NDC codes for contraception is challenging, and we relied on the list compiled by NCQA for the calculation of the HEDIS chlamydia screening measure. We have asked states to include the additional NDC codes when they calculate the measure, but to report to us the additional codes at the time of submission. This is consistent with the approach used by NCQA for the HEDIS chlamydia screening measure.

3c.2. **Describe any fees, licensing, or other requirements to use any aspect of the measure as specified** *(e.g., value/code set, risk model, programming code, algorithm)*.

Not applicable. The measure specifications, code lists, programming code and NSFG tables needed to interpret scores will all be

available at no charge on the OPA website.

4. Usability and Use

Extent to which potential audiences (e.g., consumers, purchasers, providers, policy makers) are using or could use performance results for both accountability and performance improvement to achieve the goal of high-quality, efficient healthcare for individuals or populations.

4a. Accountability and Transparency

Performance results are used in at least one accountability application within three years after initial endorsement and are publicly reported within six years after initial endorsement (or the data on performance results are available). If not in use at the time of initial endorsement, then a credible plan for implementation within the specified timeframes is provided.

4.1. **Current and Planned Use**

NQF-endorsed measures are expected to be used in at least one accountability application within 3 years and publicly reported within 6 years of initial endorsement in addition to performance improvement.

Planned	Current Use (for current use provide URL)
Public Reporting	Quality Improvement (Internal to the specific organization) Center for Medicaid and CHIP Services https://www.medicaid.gov/Medicaid-CHIP-Program-Information/By-Topics/Quality-of-Care/Maternal-and-Infant-Health-Care-Quality.html

4a.1. **For each CURRENT use, checked above, provide:**

- Name of program and sponsor
- Purpose
- Geographic area and number and percentage of accountable entities and patients included

Nearly two out of every three adult women enrolled in Medicaid are in their reproductive years (ages 19-44), and Medicaid currently finances about 45% of all births in the United States. Recognizing this, the Centers for Medicaid and CHIP Services (CMCS) has developed the Maternal and Infant Health initiative to improve the quality of maternity care and birth outcomes, and to measure how care is delivered to women. CMCS is funding 13 states to collect and report data on the new contraception measure on a developmental basis (pending NQF endorsement), to help states track the provision of postpartum contraception and to drive changes in care practices and delivery. The measure specifications developed for use in the Maternal and Infant Health initiative are found in the appendix.

4a.2. **If not currently publicly reported OR used in at least one other accountability application (e.g., payment program, certification, licensing) what are the reasons?** (*e.g., Do policies or actions of the developer/steward or accountable entities restrict access to performance results or impede implementation?*)

4a.3. **If not currently publicly reported OR used in at least one other accountability application, provide a credible plan for implementation within the expected timeframes -- any accountability application within 3 years and publicly reported within 6 years of initial endorsement.** (*Credible plan includes the specific program, purpose, intended audience, and timeline for implementing the measure within the specified timeframes. A plan for accountability applications addresses mechanisms for data aggregation and reporting.*)

The Measures Application Partnership (MAP) recommended use of the contraceptive measures in CMCS' core adult and child quality measures, if they are endorsed by NQF.

4b. Improvement

Progress toward achieving the goal of high-quality, efficient healthcare for individuals or populations is demonstrated. If not in use for performance improvement at the time of initial endorsement, then a credible rationale describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.

4b.1. **Progress on Improvement. (Not required for initial endorsement unless available.)**

Performance results on this measure (current and over time) should be provided in 1b.2 and 1b.4. Discuss:

- **Progress (trends in performance results, number and percentage of people receiving high-quality healthcare)**
- **Geographic area and number and percentage of accountable entities and patients included**

Not applicable.

4b.2. If no improvement was demonstrated, what are the reasons? If not in use for performance improvement at the time of initial endorsement, provide a credible rationale that describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.

Not applicable.

4c. Unintended Consequences

The benefits of the performance measure in facilitating progress toward achieving high-quality, efficient healthcare for individuals or populations outweigh evidence of unintended negative consequences to individuals or populations (if such evidence exists).

4c.1. Were any unintended negative consequences to individuals or populations identified during testing; OR has evidence of unintended negative consequences to individuals or populations been reported since implementation? If so, identify the negative unintended consequences and describe how benefits outweigh them or actions taken to mitigate them.

No unintended negative consequences were identified. The one issue that has arisen as a potential concern is that the measure may lead to coercive practices in which women are not offered a free choice of methods and are pressured into using the most and moderately effective, or LARC, methods (Gold 2014, Dehlendorf 2015). We do not think this will be a concern for several reasons:

Most and moderately effective methods

- Nine methods of contraception are included in the numerator for the primary measure, which are treated as being of equal value when calculating the measure. Hence, there is a wide range of client choice and the measure is not likely to prevent providers from delivering care in a fully client-centered, non-coercive, manner.

- We have deliberately not set a benchmark for the primary measure until we have more experience using the measure in real-life settings. The lack of a benchmark should lessen pressure on providers to encourage all women to use a most or moderately effective method.

LARC methods

- The focus of this measure is on ensuring access to these methods by monitoring very low rates of provision (e.g., below 2%, or well below the median of all reporting units). Further, we have explicitly proposed that this measure should not have a benchmark encouraging high rates of use, and that it would be an inappropriate measure to use in pay-for-performance or similar programs. If the measure is used as intended (i.e., to assess lack of access), this should remove pressure on providers to inappropriately "promote" LARC methods.

Both measures

- CDC-OPA recommendations describe in detail how to provide client-centered, non-coercive contraceptive counseling, and efforts to support use of the measure should be accompanied by efforts to increase awareness of the CDC-OPA recommendations (CDC/OPA 2014). Further, OPA has funded the development of training on how to provide client-centered training, which is available to all providers on the OPA-supported training website (www.fpntc.org).

References:

- Dehlendorf C, Bellanca H, Policar M (2015). Performance measures for contraceptive care: what are we actually trying to measure? Contraception. Jun;91(6):433-7.
- Gold R (2014). Guarding Against Coercion While Ensuring Access: A Delicate Balance. Guttmacher Policy Review. Summer 2014, Volume 17, Number 3.
- Harper C, Rocca CH, Thompson KM, Morfesis J, Goodman S, Darney PD, Westhoff CL, Speidel JJ (2015). Reductions in pregnancy rates in the USA with long-acting reversible contraception: a cluster randomised trial. Lancet. Volume 386, No. 9993, p562–568, 8 August 2015.
- Winner B, Peipert J, Qiuhong Z, Buckel C, Madden T et al (2012). Effectiveness of Long-Acting Reversible Contraception, The New England Journal of Medicine, 366 (21): 1998-2007.
- CDC/OPA (2014). Providing Quality Family Planning Services: Recommendations of CDC and the US Office of Population Affairs, MMWR Recommendations and Reports, April 24, 2014.
- OPA-funded training materials are available at this website: www.fpntc.org

5. Comparison to Related or Competing Measures

If a measure meets the above criteria and there are endorsed or new related measures (either the same measure focus or the same target population) or competing measures (both the same measure focus and the same target population), the measures are compared to address harmonization and/or selection of the best measure.

5. Relation to Other NQF-endorsed Measures
Are there related measures (conceptually, either same measure focus or target population) or competing measures (conceptually both the same measure focus and same target population)? If yes, list the NQF # and title of all related and/or competing measures.
Yes

5.1a. **List of related or competing measures (selected from NQF-endorsed measures)**
1517 : Prenatal & Postpartum Care (PPC)

5.1b. **If related or competing measures are not NQF endorsed please indicate measure title and steward.**

5a. Harmonization

The measure specifications are harmonized with related measures;
OR
The differences in specifications are justified

5a.1. **If this measure conceptually addresses EITHER the same measure focus OR the same target population as NQF-endorsed measure(s):**
Are the measure specifications completely harmonized?
Yes

5a.2. **If the measure specifications are not completely harmonized, identify the differences, rationale, and impact on interpretability and data collection burden.**

5b. Competing Measures

The measure is superior to competing measures (e.g., is a more valid or efficient way to measure);
OR
Multiple measures are justified.

5b.1. **If this measure conceptually addresses both the same measure focus and the same target population as NQF-endorsed measure(s):**
Describe why this measure is superior to competing measures (e.g., a more valid or efficient way to measure quality); OR provide a rationale for the additive value of endorsing an additional measure. (Provide analyses when possible.)
The proposed measure considers contraceptive care for the same population addressed in the NCQA measure on prenatal and postpartum care (PPC) (NQF#1517), although the measures address different types of services. We have aligned the contraceptive measure with the PCC measure to the extent possible, with regard to identifying the population of women with live births.

Appendix

A.1 Supplemental materials may be provided in an appendix. All supplemental materials (such as data collection instrument or methodology reports) should be organized in one file with a table of contents or bookmarks. If material pertains to a specific submission form number, that should be indicated. Requested information should be provided in the submission form and required attachments. There is no guarantee that supplemental materials will be reviewed.

Attachment **Attachment:** nqf_Appendix_Contraceptive_Care_Postpartum.pdf

Contact Information

Co.1 **Measure Steward (Intellectual Property Owner):** US Office of Population Affairs

Co.2 **Point of Contact:** Loretta, Gavin, loretta.gavin@hhs.gov, 240-453-2826-
Co.3 **Measure Developer if different from Measure Steward:** US Office of Population Affairs
Co.4 **Point of Contact:** Loretta, Gavin, loretta.gavin@hhs.gov, 240-453-2826-

Additional Information

Ad.1 **Workgroup/Expert Panel involved in measure development**
Provide a list of sponsoring organizations and workgroup/panel members' names and organizations. Describe the members' role in measure development.
The formal steward of the measure is OPA, but the measure is also supported by the U.S. Centers for Disease Control and Prevention, Division of Reproductive Health (http://www.cdc.gov/reproductivehealth/index.html).
Representatives of numerous organizations were involved in helping to develop the measure. Their roles included helping define the conceptual basis and rationale for the measure, piloting and testing data, reviewing draft measure specifications, and/or reviewing the NQF application.

- U.S. Office of Population Affairs: Susan Moskosky MS
- U.S. Centers for Disease Control and Prevention: Lisa Romero PhD, Cheryl Robbins PhD, Peter Briss MD, Karen Pazol PhD, Maria Rivera MPH
- U.S. Center for Medicaid and CHIP Services: Lekisha Daniel-Robinson
- Iowa Department of Public Health and Iowa Medicaid Enterprise: Denise Wheeler MS, Debra Kane PhD, Brittni Frederiksen PHD, Mikki Stier, Julie Lovelady, Dr. Jason Kessler, Sally Nadolsky (retired), Mark McMahon, Gerd Clabaugh, Brenda Dobson, Marcus Johnson-Miller, Abigail Holicky, Jessica Riggs
- Louisiana: Office of Outcomes Research & Evaluation, School of Pharmacy, University of Louisiana Monroe; Louisiana Department of Health & Hospitals, Bureau of Health Services Financing, Medicaid; and Louisiana Department of Health & Hospitals, Office of Public Health: Rebekah Gee MD, Beverly Hardy-Decuir PhD, Jen Steele, Ekwutosi Okoroh MD, Amy Zapata MPH, Lyn Kieltyka PhD, Angel Whittington, Eddy Myers CPA and MBA, Sandy Blake PhD, Larry Humble PharmD, and PhD
- Far Harbor LLC: Philip A. Hastings, PhD and Prasant Mohanty, MBBS, MPH
- National Contraceptive Quality Measures Workgroup

Measure Developer/Steward Updates and Ongoing Maintenance
Ad.2 **Year the measure was first released:**
Ad.3 **Month and Year of most recent revision:**
Ad.4 **What is your frequency for review/update of this measure?**
Ad.5 **When is the next scheduled review/update for this measure?**

Ad.6 **Copyright statement:** Not applicable
Ad.7 **Disclaimers:** Not applicable

Ad.8 **Additional Information/Comments:** Not applicable

MEASURE WORKSHEET

This document summarizes the evaluation of the measure as it progresses through NQF's Consensus Development Process (CDP). The information submitted by measure developers/stewards is included after the Brief Measure Information, Preliminary Analysis, and Pre-meeting Public and Member Comments sections.

To navigate the links in the worksheet: Ctrl + click link to go to the link; ALT + LEFT ARROW to return

Brief Measure Information

NQF #: 2903
Measure Title: Contraceptive Care – Most & Moderately Effective Methods
Measure Steward: US Office of Population Affairs
Brief Description of Measure: The percentage of women aged 15-44 years at risk of unintended pregnancy that is provided a most effective (i.e., sterilization, implants, intrauterine devices or systems (IUD/IUS)) or moderately effective (i.e., injectables, oral pills, patch, ring, or diaphragm) FDA-approved methods of contraception.
The proposed measure is an intermediate outcome measure because it represents a decision that is made at the end of a clinical encounter about the type of contraceptive method a woman will use, and because of the strong association between type of contraceptive method used and risk of unintended pregnancy.
Developer Rationale: Contraception is a highly effective clinical preventive service intended to prevent teen and unintended pregnancy and space births. The type of contraceptive method used by a woman is strongly associated with her risk of unintended pregnancy (Mansour 2010, Trussell 2011, Winner 2012, Secura 2014, Harper 2015). The most effective methods (LARC and sterilization) have a failure rate that is less than 1% per year under typical use (Trussell 2011). The moderately effective methods (shot, PPR, diaphragm) have a typical failure rate of 6-12% per year. And the least effective methods have a typical failure rate of 18-28%. Not using any method at all has a failure rate of 85%. The performance measures for contraceptive use are based on the fact that some methods are more effective than others at preventing unintended pregnancy, and are designed to encourage use of the most and moderately effective methods. We expect that use of this performance measure will change provider behavior in two main ways: (1) more providers will start screening women who come for non-family planning reasons about their pregnancy intention and providing them contraceptive services, as needed; and (2) when providing contraceptive services, more providers will follow ACOG, AAP and CDC recommendations to inform women about the availability of a wide range of contraceptive methods, offer client-centered education about the relative effectiveness of these methods for preventing pregnancy, as well as other aspects the client may want to consider, and take steps to ensure that a wide range of FDA-approved contraceptive methods are readily available to the client either on-site or through referral.

Sources:

- American College of Obstetricians and Gynecologists, Committee on Gynecologic Practice. Increasing access to contraceptive implants and intrauterine devices to reduce unintended pregnancy. Committee Opinion Number 642; October 2015.
- American College of Obstetricians and Gynecologists (2011). Increasing Use of Contraceptive Implants and Intrauterine Devices to Reduce Unintended Pregnancy. ACOG Committee Opinion, Number 450, December 2009 (Reaffirmed 2011).
- American College of Obstetricians and Gynecologists (2012). Adolescents and Long-Acting Reversible Contraception: Implants and Intrauterine Devices. ACOG Committee Opinion, Number 539.
- American Academy of Pediatrics (2014). Contraception for Adolescents. Pediatrics, 134:e1244–e1256.
- CDC/OPA (2014). Providing Quality Family Planning Services: Recommendations of CDC and the US Office of Population Affairs, MMWR Recommendations and Reports, April 24, 2014.
- CDC (2010). US Medical Eligibility Criteria for Contraceptive Use (USMEC), MMWR Recommendations and Reports, 59 (RR04):1–85. Available online at: http://www.cdc.gov/reproductivehealth/UnintendedPregnancy/USMEC.htm.
- Harper C, Rocca CH, Thompson KM, Morfesis J, Goodman S, Darney PD, Westhoff CL, Speidel JJ (2015). Reductions in pregnancy rates in the USA with long-acting reversible contraception: a cluster randomised trial. Lancet. Volume 386, No. 9993, p562–568, 8 August 2015
- Mansour D, Inki P, Gemzell-Danielsson K (2010). Efficacy of contraceptive methods: A review of the literature. The European Journal of Contraception and Reproductive Health Care, 15:4-16.

- Secura GM, Madden T, McNicholas C, Mullersman J, Buckel CM, Zhao Q, Peipert JF (2014). Provision of no-cost, long-acting contraception and teenage pregnancy. N Engl J Med. Oct 2;371(14):1316-23.
- Trussell J (2011). Contraceptive failure in the United States. Contraception; 83(5):397-404.
- Winner B, Peipert J, Qiuhong Z, Buckel C, Madden T et al (2012). Effectiveness of Long-Acting Reversible Contraception, The New England Journal of Medicine, 366 (21): 1998-2007.

Numerator Statement: Women aged 15-44 years of age at risk of unintended pregnancy who are provided a most (sterilization, intrauterine device, implant) or moderately (pill, patch, ring, injectable, diaphragm) effective method of contraception.
Denominator Statement: Women aged 15-44 years of age who are at risk of unintended pregnancy.
Denominator Exclusions: The following categories of women are excluded from the denominator: (1) those who are infecund for non-contraceptive reasons; (2) those who had a live birth in the last 2 months of the measurement year; or (3) those who were still pregnant or their pregnancy outcome was unknown at the end of the year.

Measure Type: Intermediate Clinical Outcome
Data Source: Administrative claims
Level of Analysis: Facility, Health Plan, Population : Regional, Population : State

IF Endorsement Maintenance – Original Endorsement Date: Most Recent Endorsement Date:

New Measure -- Preliminary Analysis

Criteria 1: Importance to Measure and Report

1a. Evidence

1a. Evidence. The evidence requirements for a *process or intermediate outcome* measure is that it is based on a systematic review (SR) and grading of the body of empirical evidence where the specific focus of the evidence matches what is being measured.

The developer provides the following evidence for this measure:

- **Systematic Review of the evidence specific to this measure?** ☒ **Yes** ☐ **No**
- **Quality, Quantity and Consistency of evidence provided?** ☒ **Yes** ☐ **No**
- **Evidence graded?** ☐ **Yes** ☒ **No**

Evidence Summary

- The developer provides a diagram outlining the steps between the structure and process that influence the intermediate outcome (use of most or moderately effective contraceptives), and how the intermediate outcome in turns influences the longer-term outcomes (reduction in unintended pregnancy).
- The measure is based on three ungraded clinical practice guidelines from the Office of Population Affairs (OPA)/CDC, ACOG, and the AAP. The guidelines recommend education and counseling for patients on contraceptive measures; two of the guidelines specifically state the counseling should begin with the most effective methods.
- Two systematic literature reviews (SR) focusing on the relationship between use of most/moderately effective contraceptives and unintended pregnancy are summarized.
- The quality of the evidence is not described in the SRs, but both are substantially comprised of randomized controlled trials.
- Results from the SRs:
  - 2011, contraceptive effectiveness: Includes estimated failure rates for a wide range of contraceptive methods under "perfect" and "typical" use. "The most effective methods (LARC and sterilization) have a failure rate that is less than 1% per year under typical use; the moderately effective methods (shot, PPR, diaphragm) have a typical failure rate of 6-12% per year; the least effective methods have a typical failure rate of 18-28%; and not using any method at all has a failure rate of 85%."

- 2010, contraceptive effectiveness: "The review broadly confirmed the hierarchy of contraceptive effectiveness in descending order as: (1) female sterilisation, long-acting hormonal contraceptives (LNG-IUS and implants); (2) Cu-IUDs with_300 mm2 surface area; (3) Cu-IUDs with5300 mm2 surface area and short-acting hormonal contraceptives (injectables, oral contraceptives, the patch and vaginal ring), and (4) barrier methods and natural methods."
- 2015, counseling effectiveness: "Five of the 8 studies that examined use of more effective methods found an increased rate of use in the intervention vs control/comparison conditions. Three studies found no significant impact. No studies found a decreased rate of use of more effective contraceptive methods."
- 2015: "Two large studies have been recently published, which provide additional evidence that: (a) the type of contraceptive method chosen is associated with risk of unintended pregnancy, and (b) that the type of counseling provided is associated with choice of method selected by the client."

Guidance from the Evidence Algorithm

Intermediate outcome measure (Box 1) → Systematic review (Box 3) → QQC present (Box 4) →SR concludes high quality evidence

Questions for the Committee:

- *What is the relationship of this measure to patient outcomes?*
- *How strong is the evidence for this relationship?*
- *Is the evidence directly applicable to the intermediate clinical outcome being measured?*

Preliminary rating for evidence: ☒ **High** ☐ **Moderate** ☐ **Low** ☐ **Insufficient**

1b. Gap in Care/Opportunity for Improvement and 1b. Disparities
Maintenance measures – increased emphasis on gap and variation

1b. Performance Gap. The performance gap requirements include demonstrating quality problems and opportunity for improvement.

- Of the 310 million people in the United States, 62 million (20%) are women of reproductive age, 15-44 years. 38 million women are at risk of unintended pregnancy and 51% of the 6.7 million pregnancies each year (3.2 million) are unintended.
- The type of contraceptive method used is strongly associated with risk of unintended pregnancy.
- The developer states "It is desirable to have a very high percentage of women at risk of unintended pregnancy using a most or moderately effective method of contraception (e.g., more than 85%)."
- The developer notes that due to limitations of claim data, "programs that serve a general population (such as a state Medicaid program) may need to take some additional steps if interested in approximating the actual percentage of women who are at risk of unintended pregnancy because claims data do not capture the following aspects needed to determine if a woman is at risk of unintended pregnancy: sexual experience, pregnancy intention, sterilization or LARC insertion in a year preceding the measurement year, and infecundity for non-contraceptive reasons (unless the woman had a procedure during the measurement year)." The developers estimate that 44% of adult women enrolled in Medicaid are not in need of contraceptive services for a variety of reasons.

	Iowa Medicaid,	Wisconsin	PPFA,	OPA Title X	OPA Title X

	ages 15-44	Medicaid, ages 15-44	ages 15-44	Clients ages 15-19	Clients ages 20-44
Number of women/sites	44,750	118,309; 17 HMOs	838,872 363 health centers	592,944	2,592,614
Provided most or moderately effective method (variation)	68% (52-67% by region, 28-79% by type of benefit)	40% (34-50% by HMO)	68% (28-90% by affiliate)	77% (34-100% by grantee)	74% (33-98% by grantee)

Disparities

- The developer states an analyses of disparities that suggests that there are statistically significant differences by age and for women who were never married. However, there were no significant differences by race/ethnicity, most categories of marital status, and poverty level (CDC/NCHS, unpublished data).

Questions for the Committee:

- *Is there a gap in care that warrants a national performance measure?*
- *Are you aware of evidence that disparities exist in this area of healthcare?*

Preliminary rating for opportunity for improvement: ☒ High ☐ Moderate ☐ Low ☐ Insufficient

Committee pre-evaluation comments

Criteria 1: Importance to Measure and Report (including 1a, 1b, 1c)

1a. Evidence to Support Measure Focus

Comments:

This is an intermediate outcome. The evidence relates directly to the intermediate outcome/desired outcomes in preventing unintended pregnancy.

**I am concerned that the evidence does not directly support this measure. The evidence has shown that the type of counseling can be associated with the choice of method selected, but measuring the provision of most or moderately effective methods does not address patient preference. A measure that addresses unintended pregnancy is important, although even this is complicated since many unintended pregnancies are desired pregnancies.

I am concerned about the focus on most or moderately effective methods is not patient-centered since it excludes other methods that might be best for a woman (ie: condoms for a woman who wants to prevent STIs or withdrawal for a patient who has used it effectively for decades). I think the measure of "providing" these contraceptive methods rather than "offering" is mistaken. Given recent data that has demonstrated strong implicit biases in contraception counseling and a history of coercion for women of color when it comes to patient autonomy, we need to be particularly careful when measuring provision as the outcome. Instead, changing this to an experience-based measure would be patient-centered (such as a measure that tracks the degree to which a patient's needs are met or a measure that tracks whether patients are screened for need/desire for contraception).**

1b. Performance Gap

Comments:

Yes, performance data on the measure was provided from Medicaid programs and national reproductive health delivery providers. The performance data indicates that there are statistically significant differences by age (i.e. younger women) and among women who were never married. However, I question the findings that there were no significant differences by race/ethnicity and poverty level. There is certainly a gap in care, it appears, between Medicaid providers and the OPA Title X providers.

85% was stated as "desired" (unclear to me how this number was obtained) but the data are significantly lower than this, so a national measure is warranted. Also complex since 44% of women do not need contraception (28% sterilized) in the data cited on page 22. Disparities definitely exist in rates of unintended pregnancy.

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability

2a1. Reliability Specifications

2a1. Specifications requires the measure, as specified, to produce consistent (reliable) and credible (valid) results about the quality of care when implemented.

Data source(s): Administrative claims

Specifications:

- The numerator is Women aged 15-44 years of age at risk of unintended pregnancy who are provided a most (sterilization, intrauterine device, implant) or moderately (pill, patch, ring, injectable, diaphragm) effective method of contraception.
- The denominator is Women aged 15-44 years of age who are at risk of unintended pregnancy.
- There are three exclusion categories: (1) those who are infecund for non-contraceptive reasons; (2) those who had a live birth in the last 2 months of the measurement year; or (3) those who were still pregnant or their pregnancy outcome was unknown at the end of the year.
- The developers tested the measure specifications based on 2014 codes, but have also included the codes needed to calculate the measure using ICD-10 and 2015 NDC codes. Both sets of codes are attached. The goal was to convert the measure to a new code set, fully consistent with the intent of the original measure.
- The measure is not risk adjusted. The developers suggest stratification by age, so that adolescents can be examined separately from adult women—this stratification is for purposes of QI but not as a method of risk adjustment.
- The calculation algorithm is included.

Questions for the Committee:

- *Are all the data elements clearly defined? Are all appropriate codes included?*
- *Is the logic or calculation algorithm clear?*
- *Is it likely this measure can be consistently implemented?*

2a2. Reliability Testing Testing attachment

2a2. Reliability testing demonstrates if the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise enough to distinguish differences in performance across providers.

SUMMARY OF TESTING

Reliability testing level ☒ **Measure score** ☐ **Data element** ☐ **Both**

Reliability testing performed with the data source and level of analysis indicated for this measure ☒ **Yes** ☐ **No**

Method(s) of reliability testing

- Testing was done using Medicaid datasets from Iowa and Louisiana and data from PPFA. Testing was performed at the facility, health plan, and public health region levels of analysis.
- The developers describe their signal-to-noise testing approach using the Intraclass Correlation (ICC) with a

Spearman-Brown prophesy formula to estimate reliability. This is an appropriate method for estimating the reliability of the measure score. Generally, reliability estimates > 0.70 are considered acceptable.

Results of reliability testing

The developers provide a table of the testing results for ICC and reliability. From the testing results, they conclude:

- "For the PPFA data both at the affiliate level and at the next level down (groups of health centers within affiliate), we found reliabilities well above the commonly accepted .90 reliability threshold for reporting and decision-making."
- "For the IME data, the rates were much more uniform by region resulting in lower ICCs, but the volume of clients still enabled adequate reliability for distinguishing performance. When region was crossed by type of health plan or benefit the contraceptive rates were more variable among the units, so even given the smaller size of these analytic units the estimated reliabilities were higher."
- The developers indicate that the large number of patients allow for strong reliability. They recommend that health centers have at least 115 patient cases for reporting rates to maintain >.70 reliability, and 450 cases to maintain >.90 reliability.

Guidance from the Reliability Algorithm

Precise specifications (Box 1)→ empirical reliability testing (Box 2) → testing of computed measure score (Box 4)→ appropriate method (Box 5) → high certainty that scores or reliable within the reporting rates recommended by the developer (Box 6a) → high

Questions for the Committee:

- *Is the test sample adequate to generalize for widespread implementation?*
- *Do the results demonstrate sufficient reliability so that differences in performance can be identified?*

Preliminary rating for reliability: ☒ **High** ☐ **Moderate** ☐ **Low** ☐ **Insufficient**

2b. Validity

2b1. Validity: Specifications

2b1. Validity Specifications. This section should determine if the measure specifications are consistent with the evidence.

Specifications consistent with evidence in 1a. ☒ **Yes** ☐ **Somewhat** ☐ **No**

Question for the Committee:

- *Are the specifications consistent with the evidence?*

2b2. Validity testing

2b2. Validity Testing should demonstrate the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality.

SUMMARY OF TESTING

Validity testing level ☒ **Measure score** ☐ **Data element testing against a gold standard** ☐ **Both**

Method of validity testing of the measure score:

- ☒ **Face validity only**
- ☐ **Empirical validity testing of the measure score**

Validity testing method:
Face validity was performed systematically by 9 experts, who were not involved in developing the measure.

Validity testing results:
The panel rated the measure on a 1-5 scale.

- The mean rating from the face validity assessment for this measure was 4.67 with a median of 5 (Strongly Agree), range 4-5. There were 66.7% (n = 6) of respondents who strongly agreed and 33.3% (n = 3) of respondents who agreed that the scores obtained from this measure, as specified, will provide an accurate reflection of quality and can be used to distinguish good and poor quality in contraceptive services.
- One respondent pointed out that "quality of the indicator will in part depend on how well 'unintended' is characterized."

Questions for the Committee:

- *Is the test sample adequate to generalize for widespread implementation?*
- *Do the results demonstrate sufficient validity so that conclusions about quality can be made?*
- *Do you agree that the score from this measure as specified is an indicator of quality?*
- *Other specific question of the validity testing?*

2b3-2b7. Threats to Validity

2b3. Exclusions:
The developer did not formally test the exclusions because "The rationale for exclusion was due to the fact that some women are not at risk of unintended pregnancy due to infecundity or pregnancy." They included the following table of exclusions:

		Number of women		
		PPFA, 2014	IME, 2013	WMP, 2014
Women 15-44 years of age		950,647	49,232	132,940
Exclusions	Infecund for non-contraceptive reasons	83	169	2,025
	Had a live birth in the last 2 months of the measurement year	7	520	2,995
	Pregnant or their pregnancy outcome was unknown at the end of the measurement year	111,685	3793	9,611
Number of women 15-44 years of age, after exclusions		838,872	44,750	118,309

Questions for the Committee:

- *Are the exclusions consistent with the evidence?*
- *Are any patients or patient groups inappropriately excluded from the measure? Are there exclusions that are missing?*
- *Are the exclusions/exceptions of sufficient frequency and variation across providers to be needed (and outweigh the data collection burden)?*

2b4. Risk adjustment: **Risk-adjustment method** ☒ **None** ☐ **Statistical model** ☐ **Stratification**

2b5. Meaningful difference *(can statistically significant and clinically/practically meaningful differences in performance*

measure scores can be identified):

- The developers provide a table of results showing the difference in performance. They state "There are very large and meaningful differences in rates across almost all levels."
- For example, the rates for most and moderately effective methods ranged by PPFA affiliate (31-86%) as well as by age; by benefit type in Iowa (from 32-74%) as well as by public health region in Iowa (31-78%), and by health plans in Wisconsin (36-45%).

Question for the Committee:

- *Does this measure identify meaningful differences about quality?*

2b6. Comparability of data sources/methods:

N/A

2b7. Missing Data

- The developers note that "The data source for this measure is claims data. Due to the nature of claims data (i.e., for billing purposes), there is typically very little missing data; further, it is difficult to ascertain when claims data is or is not missing."

Guidance from the algorithm:

Consistent with specifications (Box 1) →potential threats to validity addressed (Box 2) →face validity systematically assessed (Box 4) → substantial agreement (Box 5) → moderate (in the absence of empirical validity testing, moderate is the highest rating possible)

Preliminary rating for validity: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 2: Scientific Acceptability of Measure Properties (including all 2a, 2b, and 2d)

2a1. & 2b1. Specifications

Comments:

I'm not sure that "at risk" is defined. Does that mean any woman aged 15-44 besides the three exclusion categories? I am also unclear about why women who have had a live birth in the last 2 months of the measurement year are excluded. Can someone explain? I do not have concerns that this measure can be consistently implemented. However, I do have concerns about potential provider implicit bias or potential reproductive coercion, especially among women of color. I have read numerous studies where women of color feel targeted or coerced into using LARC more so than white women.

**Why is patient preference not accounted for in this?

I don't understand how "women at risk of unintended pregnancy" is calculated. Do they have to be sexually active? How is this decided? What about women who are started on OCPs, but then self-d/c? It's complicated to exclude women who have their IUD removed, but not women who d/c their pills.**

N/A - the specifications are consistent with evidence in 1a.

Per experts, high validity.

2a2. Reliability Testing

Comments:

Yes, reliability testing rated high. They represent sufficient reliability because of an adequate scope of entities - including geographic diversity - and patients. In addition, they used very conservative values of reliability, including medians rather than means, as well as a "floor" of reliability.

**Based on the testing, this measure is reliable.

"Is the score from this measure an indicator of quality?" Not necessarily. A patient can definitely receive exceptional counseling, and choose condoms if that is the right method for her.**

2b2. Validity Testing

Comments:
Validity testing was done using a panel that performed a face validity assessment. The panel agreed that this measure would provide an accurate reflection of quality and can be used to distinguish good and poor quality in contraceptive services. I worry that merely "providing" contraceptives rather than "offering" or even "discussing contraceptive options" may result in coercion or bias. However, the validity testing that was done is sufficient.
The evidence has looked at effectiveness of birth control methods. This looks at provision of birth control methods, so the previous evidence doesn't validate this test.

2b3. Exclusions Analysis
2b4. Risk Adjustment/Stratification for Outcome or Resource Use Measures
2b5. Identification of Statistically Significant & Meaningful Differences In Performance
2b6. Comparability of Performance Scores When More Than One Set of Specifications
2b7. Missing Data Analysis and Minimizing Bias
Comments:
**- All threats were examined. Since this measure deals with claims data there is very little potential for threats.
- Exclusions seem consistent (except, again, I am curious why women who were pregnant within two months were excluded).
- No risk adjustment was justified because the potential variations in contraceptive use are based on systemic - clinical and programmatic - considerations.
- This measure identifies meaningful differences in rates across along all levels (age group, rate, level), and reinforces that there is room for improvement.
- This is claims data so there is very little missing data and, even so, the missing data does not constitute a threat to the validity of this measure.**
Significant differences in performance rates is concerning for validity, but might also indicate that some clinics really need a performance measure to improve. The exclusions seem reasonable to me. Missing data will be problematic.

Criterion 3. Feasibility
Maintenance measures – no change in emphasis – implementation issues may be more prominent

3. Feasibility is the extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

- This measure is based on administrative claims. Claims data is generally thought to be of minimal burden and highly feasible.
- The measure has been piloted by Planned Parenthood and two state Medicaid programs.

Questions for the Committee:
- *Are the required data elements routinely generated and used during care delivery?*
- *Are the required data elements available in electronic form, e.g., EHR or other electronic sources?*
- *Is the data collection strategy ready to be put into operational use?*

Preliminary rating for feasibility: ☒ **High** ☐ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments
Criteria 3: Feasibility

3a. Byproduct of Care Processes
3b. Electronic Sources
3c. Data Collection Strategy

Comments:
**Collecting the data elements is routinely generated during care delivery, so I do not have any concerns about that. It does not represent an undue burden, in my perspective, to collect this data. Though, as measure developers noted, there is a limitation in estimating unintended pregnancy. **
Very feasible as is, although I would argue that we should change the measure to look at whether contraceptive methods were offered, rather than provided, which would be harder to analyze.

Criterion 4: Usability and Use
Maintenance measures – increased emphasis – much greater focus on measure use and usefulness, including both impact /improvement and unintended consequences

4. Usability and Use evaluate the extent to which audiences (e.g., consumers, purchasers, providers, policymakers) use or could use performance results for both accountability and performance improvement activities.

Current uses of the measure
Publicly reported? ☐ **Yes** ☒ **No**

Current use in an accountability program? ☐ **Yes** ☐ **No**
OR
Planned use in an accountability program? ☒ **Yes** ☐ **No**

Accountability program details

- Public reporting planned
- Centers for Medicaid and CHIP Services (CMCS) has developed the Maternal and Infant Health initiative to improve the quality of maternity care and birth outcomes, and to measure how care is delivered to women. CMCS is funding 13 states to collect and report data on the new contraception measure on a developmental basis (pending NQF endorsement), to help states track the provision of postpartum contraception and to drive changes in care practices and delivery.
- For internal quality improvement: Planned Parenthood Federation of America a Clinical Quality Improvement (CQI) Department (70% of affiliates)
- OPA Title X family planning program plans to publically report on the performance of all grantees within three years: www.familyplanningdashboard.com

Improvement results NA

Unexpected findings (positive or negative) during implementation
None identified

Potential harms
The developers point out "The one issue that has arisen as a potential concern is that the measure may lead to coercive practices in which women are not offered a free choice of methods and are pressured into using the most and moderately effective, or LARC, methods (Gold 2014, Dehlendorf 2015)." The developers provide an explanation of why they do not believe this will be a significant issue in using this measure

Feedback:
The Measures Application Partnership (MAP) recommended use of the contraceptive measures in CMCS' core adult and child quality measures, if they are endorsed by NQF.

***Questions for the Committee**:*

- *How can the performance results be used to further the goal of high-quality, efficient healthcare?*
- *Do the benefits of the measure outweigh any potential unintended consequences?*

Preliminary rating for usability and use: ☒ **High** ☐ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 4: Usability and Use

4a. Accountability and Transparency
4b. Improvement
4c. Unintended Consequences

Comments:

Measure developers mention the concern about coercive practices in contraceptive decision-making. Measure developers respond to these concerns, which I think are sufficient responses.
High concern for unintended consequence of coercion in birth control options counseling. It's important to help move payers to improve access of all birth control options to patients, but providers need to not feel pressured to increase LARC uptake. Might be helpful changing the standard to a minimum - all clinics must provide at least 1% of their clients with LARC, to make sure that LARC services are available.

Criterion 5: Related and Competing Measures

Related or competing measures

- 2902: Contraceptive Care – Postpartum a
- 2904: Access to LARC

Harmonization

- Measures 2903,2902 and 2904 are from the same developer and harmonized.

Pre-meeting public and member comments

- Planned Parenthood Federation of America, the nation's leading provider of women's reproductive healthcare, supports the endorsement of the proposed measures. Contraception is an important and effective preventive service to reduce unintended pregnancy as well as improve birth spacing and family planning. PPFA provided de-identified data included in the application to demonstrate the reliability and validity of the measures as well the feasibility of using them for quality improvement. Currently, PPFA has already begun using a developmental version of these measures for quality improvement and looks forward to incorporate NQF endorsed measure into its portfolio of internal quality improvement work. National endorsement of these new performance measures on contraceptive care aligns with the April 2015 call by the Institute of Medicine for standardized metrics that include measuring contraceptive use to support reducing unintended pregnancy. Further, these will be the first nationally endorsed measures on contraceptive care, providing important tools to all providers who serve women of reproductive age.

NATIONAL QUALITY FORUM—Evidence (subcriterion 1a)

Measure Number (*if previously endorsed*)**:** Click here to enter NQF number
Measure Title Contraceptive Care –Most and Moderately Effective Methods
IF the measure is a component in a composite performance measure, provide the title of the Composite Measure here: Click here to enter composite measure #/ title

Date of Submission: 2/15/2016

Instructions

- *For composite performance measures:*
  - *A separate evidence form is required for each component measure unless several components were studied together.*
  - *If a component measure is submitted as an individual performance measure, attach the evidence form to the individual measure submission.*
- Respond to all questions as instructed with answers immediately following the question. All information needed to demonstrate meeting the evidence subcriterion (1a) must be in this form. An appendix of *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 10 pages (*incudes questions/instructions*; minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.

Note: The information provided in this form is intended to aid the Steering Committee and other stakeholders in understanding to what degree the evidence for this measure meets NQF's evaluation criteria.

1a. Evidence to Support the Measure Focus
The measure focus is evidence-based, demonstrated as follows:

- Health outcome: [3] a rationale supports the relationship of the health outcome to processes or structures of care. Applies to patient-reported outcomes (PRO), including health-related quality of life/functional status, symptom/symptom burden, experience with care, health-related behavior.
- Intermediate clinical outcome: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured intermediate clinical outcome leads to a desired health outcome.
- Process: [5] a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured process leads to a desired health outcome.
- Structure: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured structure leads to a desired health outcome.
- Efficiency: [6] evidence not required for the resource use component.

Notes

3. Generally, rare event outcomes do not provide adequate information for improvement or discrimination; however, serious reportable events that are compared to zero are appropriate outcomes for public reporting and quality improvement.
4. The preferred systems for grading the evidence are the U.S. Preventive Services Task Force (USPSTF) grading definitions and methods, or Grading of Recommendations, Assessment, Development and Evaluation (GRADE) guidelines.
5. Clinical care processes typically include multiple steps: assess → identify problem/potential problem → choose/plan intervention (with patient input) → provide intervention → evaluate impact on health status. If the measure focus is one step in such a multistep process, the step with the strongest evidence for the link to the desired outcome should be selected as the focus of measurement. Note: A measure focused only on collecting PROM data is not a PRO-PM.
6. Measures of efficiency combine the concepts of resource use and quality (see NQF's Measurement Framework: Evaluating Efficiency Across Episodes of Care; AQA Principles of Efficiency Measures).

1a.1.This is a measure of: (*should be consistent with type of measure entered in De.1*)

Outcome

☐ Health outcome: Click here to name the health outcome

☐Patient-reported outcome (PRO): Click here to name the PRO

PROs include HRQoL/functional status, symptom/symptom burden, experience with care, health-related behaviors

☒ Intermediate clinical outcome (*e.g., lab value*): Provision of most/moderately effective methods of contraception

☐ Process: Click here to name the process

☐ Structure: Click here to name the structure

☐ Other: Click here to name what is being measured

_____

HEALTH OUTCOME/PRO PERFORMANCE MEASURE *If not a health outcome or PRO, skip to 1a.3*

1a.2. Briefly state or diagram the path between the health outcome (or PRO) and the healthcare structures, processes, interventions, or services that influence it.

Not a health outcome or PRO.

1a.2.1. State the rationale supporting the relationship between the health outcome (or PRO) to at least one healthcare structure, process, intervention, or service (*i.e., influence on outcome/PRO*).

Note: For health outcome/PRO performance measures, no further information is required; however, you may provide evidence for any of the structures, processes, interventions, or service identified above.

_____

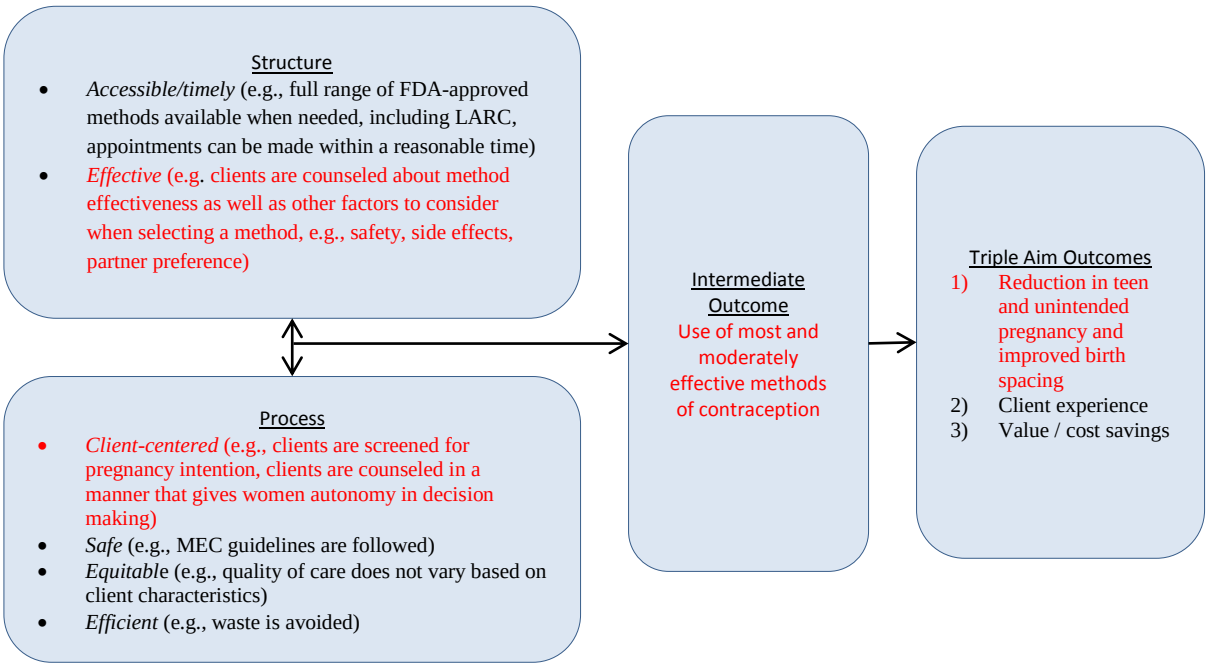
INTERMEDIATE OUTCOME, PROCESS, OR STRUCTURE PERFORMANCE MEASURE

1a.3. Briefly state or diagram the path between structure, process, intermediate outcome, and health outcomes. Include all the steps between the measure focus and the health outcome.

The diagram below illustrates the steps between the structure and process that influence the intermediate outcome, and how the intermediate outcome in turns influences the longer-term outcomes. The text highlighted in red shows the primary relationships that will be affected by use of the proposed measure: (a) increased use of the most and moderately effective methods of contraception will influence rates of unintended pregnancy; and (b) appropriate counseling of a client can lead to increased use of the most and moderately effective methods of contraception.

The type of contraceptive method used by a woman is strongly associated with her risk of unintended pregnancy. The most effective methods (sterilization and the long-acting reversible contraceptive [LARC] methods of intrauterine devices and implants) have a failure rate that is less than 1% per year under typical use; the moderately effective methods (shot, oral pills, patch, ring, and diaphragm) have a typical failure rate of 6-12% per year; the least effective methods have a typical failure rate of 18-28%; and if no method is used then 85 of every 100 women will become pregnant in a year (Trussell 2011).

The measure is secondarily supported by evidence that the way in which contraceptive counseling is offered (e.g., increased screening of clients for reproductive intention; the provision of client-centered counseling, which includes providing information about and ready access to the most and moderately effective methods of contraception; and ready access to all methods of contraception, ideally on a same-day basis) will lead to increased use of the most and moderately effective methods of contraception (i.e., the intermediate outcome).



1a.3.1. What is the source of the systematic review of the body of evidence that supports the performance measure?

☒ Clinical Practice Guideline recommendation – ***complete sections*** 1a.4***, and*** 1a.7

☐ US Preventive Services Task Force Recommendation – ***complete sections*** 1a.5 ***and*** 1a.7

☒ Other systematic review and grading of the body of evidence (*e.g., Cochrane Collaboration, AHRQ Evidence Practice Center*) – ***complete sections*** 1a.6 ***and*** 1a.7

☐ Other – ***complete section*** 1a.8

Please complete the sections indicated above for the source of evidence. You may skip the sections that do not apply.

1a.4. CLINICAL PRACTICE GUIDELINE RECOMMENDATION

1a.4.1. Guideline citation (*including date*) and **URL for guideline** (*if available online*):

Clinical recommendations (from both government sources and professional organizations) are the best source of evidence about the relationship between contraceptive counseling and increased use of the most and moderately effective methods of contraception (see diagram above).

CDC/OPA (2014). Providing Quality Family Planning Services (QFP): Recommendations of CDC and the US Office of Population Affairs, MMWR Recommendations and Reports, April 24, 2014. http://www.cdc.gov/mmwr/preview/mmwrhtml/rr6304a1.htm

American College of Obstetricians and Gynecologists (ACOG), Committee on Gynecologic Practice. Increasing access to contraceptive implants and intrauterine devices to reduce unintended pregnancy. Committee Opinion Number 642; October 2015.

The American Academy of Pediatrics (AAP) (2014). Contraception for Adolescents. Pediatrics, 134:e1244–e1256.

1a.4.2. Identify guideline recommendation number and/or page number and **quote verbatim, the specific guideline recommendation**.

"Providers are encouraged to present information on potential reversible methods of contraception by using a tiered approach (i.e., presenting information on the most effective methods first, before presenting information on less effective methods). This information should include an explanation that long-acting reversible contraceptive methods are safe and effective for most women, including those who have never given birth and adolescents. Information should be tailored and presented to ensure a client-centered approach. It is not appropriate to omit presenting information on a method solely because the method is not available at the service site. If not all methods are available at the service site, it is important to have strong referral links in place to other providers to maximize opportunities for clients to obtain their preferred method that is medically appropriate."
Source: CDC/OPA (2014). *Providing Quality Family Planning Services*, page 8 and Appendix B

"For all women at risk of unintended pregnancy, obstetrician-gynecologists should provide counseling on all contraceptive options, including implants and IUDs. Long-acting reversible contraception methods require a single action of motivation for long-term use, eliminating adherence and user dependence from the effectiveness equation. These top-tier methods share the highest continuation rates of all contraceptives, which is one of the most important factors in contraceptive success."
Source: ACOG (2015), page 1.

"Contraceptive methods most commonly used by adolescents are listed below, ordered from most to least effective, starting with long-acting reversible contraception (LARC); implants and IUDs. *Pediatricians are encouraged to counsel adolescents in that order, discussing the most effective contraceptive methods first*." ACOG (2014), page e1246.

1a.4.3. Grade assigned to the quoted recommendation <u>with definition</u> of the grade:

Not applicable

1a.4.4. Provide all other grades and associated definitions for recommendations in the grading system. (*Note: If separate grades for the strength of the evidence, report them in section 1a.7.*)

Not applicable

1a.4.5. Citation and URL for methodology for grading recommendations (*if different from 1a.4.1*)**:**

Not applicable

1a.4.6. If guideline is evidence-based (rather than expert opinion), are the details of the quantity, quality, and consistency of the body of evidence available (e.g., evidence tables)?

☒ Yes → ***complete section*** 1a.7
☐ No → report on another systematic review of the evidence in sections 1a.6 and 1a.7; if another review does not exist, ***provide what is known from the guideline review of evidence in*** 1a.7

1a.5. UNITED STATES PREVENTIVE SERVICES TASK FORCE RECOMMENDATION
1a.5.1. Recommendation citation (*including date*) and **URL for recommendation** (*if available online*):

Not applicable

1a.5.2. Identify recommendation number and/or page number and **quote verbatim, the specific recommendation**.

Not applicable

1a.5.3. Grade assigned to the quoted recommendation with definition of the grade:

Not applicable

1a.5.4. Provide all other grades and associated definitions for recommendations in the grading system. (*Note: the grading system for the evidence should be reported in section 1a.7.*)

Not applicable

1a.5.5. Citation and URL for methodology for grading recommendations (*if different from 1a.5.1*)**:**

Not applicable

Complete section 1a.7

_____

1a.6. OTHER SYSTEMATIC REVIEW OF THE BODY OF EVIDENCE

1a.6.1. Citation (*including date*) and **URL** (*if available online*):

Two systematic literature reviews are the best source of evidence about the relationship between use of most and moderately effective methods of contraception and unintended pregnancy (see diagram in 1a.3, above).

The first review was led by Professor James Trussell from Princeton University, which is repeated on an ongoing basis and published in a handbook entitled "Contraceptive Technology". The Trussell analyses serve as the primary source of information about contraceptive failure rates, and are cited by the World Health Organization, CDC, and leading professional associations in the U.S. and in other countries. Trussell used two sources of data when estimating contraceptive failure. The first was published research, which comprised results from clinical trials and surveys. The second source was the CDC's National Survey of Family Growth (NSFG), which was used to estimate *typical* use rates using data from a nationally representative sample of users.

- Trussell J (2011). Contraceptive efficacy. In: Hatcher RA, Trussell J, Nelson AL, Cates W, Kowal D, editors. Contraceptive technology: twentieth revised edition. New York: Ardent Media; 2011, pp. 777–861. This was subsequently summarized in: Trussell J (2011). Contraceptive failure in the United States. Contraception; 83(5):397-404.

- WHO/Department of Reproductive Health and Research & Johns Hopkins Bloomberg School of Public Health/Center for Communication Programs (2011). Family Planning: A Global Handbook for Providers. Baltimore and Geneva: CCP and WHO.

The second review was conducted by Mansour et al in 2010. They search Medline and Embase from January 1990 to February 2008 for publications reporting contraceptive failure rates.

- Mansour D, Inki P, Gemzell-Danielsson K (2010). Efficacy of contraceptive methods: A review of the literature. The European Journal of Contraception and Reproductive Health Care, 15:4-16.

1a.6.2. Citation and URL for methodology for evidence review and grading (*if different from 1a.6.1*)**:**

Evidence of contraceptive effectiveness & its impact on unintended pregnancy

- See 1a.6.1 above

Evidence of effectiveness of counseling or other interventions to affect patients' choice of method

- Zapata LB, Tregear SJ, Curtis KM, Tiller M, Pazol K, Mautone-Smith N, Gavin LE (2015). Impact of Contraceptive Counseling in Clinical Settings: A Systematic Review. Am J Prev Med. 2015 Aug;49(2 Suppl 1):S31-45.

Complete section 1a.7

_____

1a.7. FINDINGS FROM SYSTEMATIC REVIEW OF BODY OF THE EVIDENCE SUPPORTING THE MEASURE

If more than one systematic review of the evidence is identified above, you may choose to summarize the one (or more) for which the best information is available to provide a summary of the quantity, quality, and consistency of the body of evidence. Be sure to identify which review is the basis of the responses in this section and if more than one, provide a separate response for each review.

1a.7.1. What was the specific structure, treatment, intervention, service, or intermediate outcome addressed in the evidence review?

Evidence of contraceptive effectiveness & its impact on unintended pregnancy

The studies examining contraceptive efficacy and effectiveness considered the impact of use of specific contraceptive methods on risk of pregnancy (i.e., contraceptive failure). Pregnancy risk can be assessed either through life table analyses (usually through 12 months) that show the percentage of women who become pregnant, or the score on the Pearl Index. The Pearl Index is a commonly used technique for reporting the effectiveness of a birth control method in clinical trials, and estimates the number of unintended pregnancies over a period of exposure (e.g. 100 women over one year of use, or 10 women over 10 years). Contraceptive failure rates are reported for *perfect use* and *typical use*. Perfect use reflects how effective methods can be in preventing pregnancy when used consistently and correctly according to instructions. Typical use reflects how effective methods are for the average person who does not always use methods correctly or consistently. Pregnancy rates during typical use of adherence-dependent methods (such as the oral pill) generally vary widely for different groups using the same method, primarily due to differences in the propensity to use the method perfectly.

Evidence of effectiveness of counseling or other interventions to affect patients' choice of method

The systematic review underpinning the CDC-OPA recommendation on contraceptive counseling used an analytic framework that considered the impact of providing contraceptive counseling and/or education on short (e.g., client knowledge, attitudes), medium (e.g., selection of more effective methods, correct and consistent use) and long-term (unintended pregnancy) outcomes (Zapata 2015).

1a.7.2. Grade assigned for the quality of the quoted evidence with definition of the grade:

Evidence of contraceptive effectiveness & its impact on unintended pregnancy

While the quality of the studies was not graded in either the Trussell (2011) or Mansour (2010) review, it was significantly based on systematic reviews of randomized controlled trials.

Evidence of effectiveness of counseling or other interventions to affect patients' choice of method

The review did not grade the *overall* body of evidence. However, the quality of individual studies was graded in accordance with USPSTF methodologies for doing so, i.e., Level I, Level II-1, Level II-2, Level II-3, Level III.

1a.7.3. Provide all other grades and associated definitions for strength of the evidence in the grading system.

Not applicable

1a.7.4. What is the time period covered by the body of evidence? (*provide the date range, e.g., 1990-2010*). Date range

Trussell (2011): 1958-2010
Mansour (2010): January 1990 to February 2008
Zapata (2015): 1985-February 2011 with supplemental searches through 2014

QUANTITY AND QUALITY OF BODY OF EVIDENCE

1a.7.5. How many and what type of study designs are included in the body of evidence? (*e.g., 3 randomized controlled trials and 1 observational study*)

Evidence of contraceptive effectiveness & its impact on unintended pregnancy

- Trussell et al 2011: The review comprised results from clinical trials and surveys; the most recent review listed more than 350 studies, of which the majority was randomized controlled trials (Trussell 2011a).

- Mansour et al 2010: The authors identified and extracted information from 139 publications. Of the included studies, 47 assessed combined oral contraceptives (COCs), one assessed progestogen-only pills (POPs), three assessed the patch, three assessed the vaginal ring, 15 assessed implants, 16 assessed injectables, 31 assessed copper intrauterine devices (Cu-IUDs), nine assessed the levonorgestrel-releasing intrauterine system (LNGIUS), three assessed the male condom, four assessed other barrier methods, 11 assessed natural methods, and four assessed female sterilization. Overall, there were 64 publications of randomized controlled studies included in this review. A detailed description of each publication can be accessed from www.informahealthcare.com/doi/pdf/10.3109/13625180903427675.

Evidence of effectiveness of counseling or other interventions to affect patients' choice of method

  - Zapata et al (2015): 22 studies (from 23 articles) met the inclusion criteria; 8 studies included use of more effective methods as an outcome. Seven of the 8 studies were randomized controlled trials, while the eighth utilized a pre-posttest study design.

1a.7.6. What is the overall quality of evidence across studies in the body of evidence? (*discuss the certainty or confidence in the estimates of effect particularly in relation to study factors such as design flaws, imprecision due to small numbers, indirectness of studies to the measure focus or target population*)

The quality of evidence is not described in either the Trussell (2011) or the Mansour (2010) publications. However, both reviews are substantially comprised of randomized controlled trials.

In Zapata et al (2011), 7 of the 8 studies were graded Level I (properly designed randomized controlled trial), and the 8th study was graded Level II-3 (evidence obtained from time series, uncontrolled trial).

ESTIMATES OF BENEFIT AND CONSISTENCY ACROSS STUDIES IN BODY OF EVIDENCE

1a.7.7. What are the estimates of benefit—magnitude and direction of effect on outcome(s) across studies in the body of evidence? (*e.g., ranges of percentages or odds ratios for improvement/ decline across studies, results of meta-analysis, and statistical significance*)

Evidence of contraceptive effectiveness & its impact on unintended pregnancy

- Trussell et al 2011: The key findings of this review are estimated failure rates for a wide range of contraceptive methods under "perfect" and "typical" use. The most recent findings – published in 2011 -- are that the most effective methods (LARC and sterilization) have a failure rate that is less than 1% per year under typical use; the moderately effective methods (shot, PPR, diaphragm) have a typical failure rate of 6-12% per year; the least effective methods have a typical failure rate of 18-28%; and not using any method at all has a failure rate of 85%.

- Mansour et al 2010: "Information was identified and extracted from 139 studies. One-year Pearl Indices reported for short-acting user-dependent hormonal methods were generally less than 2.5. Gross life-table rates for long-acting hormonal methods (implants and the levonorgestrel releasing-intrauterine system [LNG-IUS]) generally ranged between 0–0.6 per 100 at one year, but wider ranges (0.1–1.5 per 100) were observed for the copper intrauterine devices (0.1–1.4 per 100 for Cu-IUDs with surface area _300 mm2 and 0.6–1.5 per 100 for those with surface area5300 mm2). Barrier and natural methods were the least effective." The authors conclude that "the review broadly confirmed the hierarchy of contraceptive effectiveness in descending order as: (1) female sterilisation, long-acting hormonal contraceptives (LNG-IUS and implants); (2) Cu-IUDs with_300 mm2 surface area; (3) Cu-IUDs with5300 mm2 surface area and short-acting hormonal contraceptives (injectables, oral contraceptives, the patch and vaginal ring), and (4) barrier methods and natural methods."

Evidence of effectiveness of counseling or other interventions to affect patients' choice of method

- Zapata (2015): Five of the 8 studies that examined use of more effective methods found an increased rate of use in the intervention vs control/comparison conditions. Three studies found no significant impact. No studies found a decreased rate of use of more effective contraceptive methods.

1a.7.8. What harms were studied and how do they affect the net benefit (benefits over harms)?

The harms were not noted in the cited reviews. However, CDC clinical recommendations on contraceptive safety explicitly address this question. CDC's "US Medical Eligibility Criteria for Contraceptive Use" (USMEC) describe what contraceptive methods are safe for women with a range of characteristics (e.g., age, postpartum) and medical conditions (e.g., infectious or chronic diseases). The citation for the USMEC recommendations is:

> CDC (2010). US Medical Eligibility Criteria for Contraceptive Use, MMWR Recommendations and Reports, 59 (RR04):1–85. Available online at: http://www.cdc.gov/reproductivehealth/UnintendedPregnancy/USMEC.htm."

The evidence on which the USMEC recommendations are based has been summarized in the following journal supplement:

> Contraception, Volume 82, Issue 1, Pages 1-118 (July 2010). Available online at: http://www.sciencedirect.com/science/journal/00107824/82/1

UPDATE TO THE SYSTEMATIC REVIEW(S) OF THE BODY OF EVIDENCE

1a.7.9. If new studies have been conducted since the systematic review of the body of evidence, provide for each new study: 1) citation, 2) description, 3) results, 4) impact on conclusions of systematic review.

Results from two large studies have been recently published, which provide additional evidence that: (a) the type of contraceptive method chosen is associated with risk of unintended pregnancy, and (b) that the type of counseling provided is associated with choice of method selected by the client. The first study is a cluster-randomized trial led by

researchers at the University of California – San Francisco (Harper 2015) and the second is a prospective cohort study that is known as "Project CHOICE" (Winner 2012).

UCSF trial (Harper et al 2015)

A cluster randomized trial was conducted in 2011-2013 to assess the effects of an intervention to increase patients' access to long-acting reversible contraceptives (LARCs) on pregnancy rates. A total of 40 clinics participated: 20 clinics were randomly assigned to receive evidence-based training on providing counselling and insertion of intrauterine devices (IUDs) or progestin implants, and 20 to provide standard care. Usual costs for contraception were maintained at all sites. Women aged 18-25 years attending family planning or abortion care visits and not desiring pregnancy in the next 12 months were recruited. The primary outcome was selection of an IUD or implant at the clinic visit and secondary outcome was pregnancy within 12 months. Generalised estimating equations for clustered data were used to measure the intervention effect on contraceptive selection, and survival analysis was used to assess pregnancy rates. Of 1500 women enrolled, more at intervention than control sites reported receiving counselling on IUDs or implants (565 [71%] of 797 vs 271 [39%] of 693, odds ratio 3·8, 95% CI 2·8-5·2) and more selected LARCs during the clinic visit (224 [28%] vs 117 [17%], 1·9, 1·3-2·8). The pregnancy rate was lower in intervention group than in the control group after family planning visits (7·9 vs 15·4 per 100 person-years), but not after abortion visits (26·5 vs 22·3 per 100 person-years). We found a significant intervention effect on pregnancy rates in women attending family planning visits (hazard ratio 0·54, 95% CI 0·34-0·85).

- Harper C, Rocca CH, Thompson KM, Morfesis J, Goodman S, Darney PD, Westhoff CL, Speidel JJ (2015). Reductions in pregnancy rates in the USA with long-acting reversible contraception: a cluster randomised trial. Lancet. Volume 386, No. 9993, p562–568, 8 August 2015

Project CHOICE (Secura et al 2014, Winner et al 2015)

The Contraceptive CHOICE Project was a prospective cohort study involving 9256 St. Louis area adolescent and adult women 14 to 45 years of age, in which women were counseled about the use of LARC methods to prevent unintended pregnancy. Participants were educated about reversible contraception, with an emphasis on the benefits of LARC methods, were provided with their choice of reversible contraception at no cost, and were followed for 2 to 3 years. Almost three-quarters of enrolled participants chose a LARC method when they were counseled about effectiveness and offered their choice of method at no charge, and continuation rates were high 2 years (77% for LARC users vs 41% for non-LARC users) and 3 years (67% for LARC users vs 31% for non-LARC users) after insertion. The contraceptive failure rate among participants using pills, patch, or ring was 4.55 per 100 participant-years, as compared with 0.27 among participants using long-acting reversible contraception (hazard ratio after adjustment for age, educational level, and history with respect to unintended pregnancy, 21.8; 95% confidence interval, 13.7 to 34.9).

- Winner B, Peipert J, Qiuhong Z, Buckel C, Madden T et al (2012). Effectiveness of Long-Acting Reversible Contraception, The New England Journal of Medicine, 366 (21): 1998-2007
- Diedrich, J.T., et al., *Three-year continuation of reversible contraception.* Am J Obstet Gynecol, 2015. **213**(5): p. 662 e1-8.
- O'Neil-Callahan, M., et al., *Twenty-four-month continuation of reversible contraception.* Obstet Gynecol, 2013. **122**(5): p. 1083-91.

_____

1a.8 OTHER SOURCE OF EVIDENCE

If source of evidence is NOT from a clinical practice guideline, USPSTF, or systematic review, please describe the evidence on which you are basing the performance measure.

Not applicable.

1a.8.1 What process was used to identify the evidence?

Not applicable.

1a.8.2. Provide the citation and summary for each piece of evidence.

Not applicable.

1. Evidence, Performance Gap, Priority – Importance to Measure and Report

Extent to which the specific measure focus is evidence-based, important to making significant gains in healthcare quality, and improving health outcomes for a specific high-priority (high-impact) aspect of healthcare where there is variation in or overall less-than-optimal performance. ***Measures must be judged to meet all subcriteria to pass this criterion and be evaluated against the remaining criteria.***

1a. Evidence to Support the Measure Focus – See attached Evidence Submission Form
nqf_evidence_Contraceptive_Care_MOST_MOD.docx

1b. Performance Gap

Demonstration of quality problems and opportunity for improvement, i.e., data demonstrating:

- considerable variation, or overall less-than-optimal performance, in the quality of care across providers; and/or
- disparities in care across population groups.

1b.1. **Briefly explain the rationale for this measure** *(e.g., the benefits or improvements in quality envisioned by use of this measure)*
Contraception is a highly effective clinical preventive service intended to prevent teen and unintended pregnancy and space births. The type of contraceptive method used by a woman is strongly associated with her risk of unintended pregnancy (Mansour 2010, Trussell 2011, Winner 2012, Secura 2014, Harper 2015). The most effective methods (LARC and sterilization) have a failure rate that is less than 1% per year under typical use (Trussell 2011). The moderately effective methods (shot, PPR, diaphragm) have a typical failure rate of 6-12% per year. And the least effective methods have a typical failure rate of 18-28%. Not using any method at all has a failure rate of 85%. The performance measures for contraceptive use are based on the fact that some methods are more effective than others at preventing unintended pregnancy, and are designed to encourage use of the most and moderately effective methods. We expect that use of this performance measure will change provider behavior in two main ways: (1) more providers will start screening women who come for non-family planning reasons about their pregnancy intention and providing them contraceptive services, as needed; and (2) when providing contraceptive services, more providers will follow ACOG, AAP and CDC recommendations to inform women about the availability of a wide range of contraceptive methods, offer client-centered education about the relative effectiveness of these methods for preventing pregnancy, as well as other aspects the client may want to consider, and take steps to ensure that a wide range of FDA-approved contraceptive methods are readily available to the client either on-site or through referral.

Sources:

- American College of Obstetricians and Gynecologists, Committee on Gynecologic Practice. Increasing access to contraceptive implants and intrauterine devices to reduce unintended pregnancy. Committee Opinion Number 642; October 2015.
- American College of Obstetricians and Gynecologists (2011). Increasing Use of Contraceptive Implants and Intrauterine Devices to Reduce Unintended Pregnancy. ACOG Committee Opinion, Number 450, December 2009 (Reaffirmed 2011).
- American College of Obstetricians and Gynecologists (2012). Adolescents and Long-Acting Reversible Contraception: Implants and Intrauterine Devices. ACOG Committee Opinion, Number 539.
- American Academy of Pediatrics (2014). Contraception for Adolescents. Pediatrics, 134:e1244–e1256.
- CDC/OPA (2014). Providing Quality Family Planning Services: Recommendations of CDC and the US Office of Population Affairs, MMWR Recommendations and Reports, April 24, 2014.
- CDC (2010). US Medical Eligibility Criteria for Contraceptive Use (USMEC), MMWR Recommendations and Reports, 59 (RR04):1–85. Available online at: http://www.cdc.gov/reproductivehealth/UnintendedPregnancy/USMEC.htm.
- Harper C, Rocca CH, Thompson KM, Morfesis J, Goodman S, Darney PD, Westhoff CL, Speidel JJ (2015). Reductions in pregnancy rates in the USA with long-acting reversible contraception: a cluster randomised trial. Lancet. Volume 386, No. 9993, p562–568, 8 August 2015
- Mansour D, Inki P, Gemzell-Danielsson K (2010). Efficacy of contraceptive methods: A review of the literature. The European Journal of Contraception and Reproductive Health Care, 15:4-16.
- Secura GM, Madden T, McNicholas C, Mullersman J, Buckel CM, Zhao Q, Peipert JF (2014). Provision of no-cost, long-acting contraception and teenage pregnancy. N Engl J Med. Oct 2;371(14):1316-23.
- Trussell J (2011). Contraceptive failure in the United States. Contraception; 83(5):397-404.
- Winner B, Peipert J, Qiuhong Z, Buckel C, Madden T et al (2012). Effectiveness of Long-Acting Reversible Contraception, The New England Journal of Medicine, 366 (21): 1998-2007.

1b.2. **Provide performance scores on the measure as specified (current and over time) at the specified level of analysis.** *(This is*

required for endorsement maintenance. Include mean, std dev, min, max, interquartile range, scores by decile. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included). This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.
Performance scores for the contraceptive care measure are presented for four programs; two are state Medicaid programs (i.e., the Iowa Medicaid Enterprise and the Wisconsin Medicaid Program) and two ae organizations that focus on the delivery of reproductive health services (i.e., the Planned Parenthood Federation of America, and OPA's Title X family planning program).

Iowa's Medicaid Enterprise (IME) provides contraceptive services to women through its general Medicaid program and its family planning waiver program (IFPN). Services are available to Iowa residents who are US citizens or qualified immigrants. To be eligible for IFPN services, the following guidelines apply: an individual does not have insurance or the insurance does not cover family planning services; the individual is a man or woman between the ages of 12 and 54; and family income is at or below 300 percent of the federal poverty level. In addition, women whose pregnancy and delivery was covered by Medicaid also qualify automatically for family planning services. A total of 44,750 women met the inclusion criteria and were included in the analysis. The results showed that 68% of clients ages 15-44 were provided a most or moderately effective method of contraception; there was variation by public health region (52 to 67%) and type of benefit (i.e., general Medicaid vs IFPN family planning waiver) (28 to 79%). For more details, see the attached testing report.

The Wisconsin Medicaid Program (WMP) provides contraceptive services to women through its general Medicaid (BadgerCare Plus) and family planning only services (FPOS) programs. Services are available to Wisconsin residents who are US citizens or qualified immigrants meeting income eligibility criteria (e.g., a child <18 years with household income at or below 300% FPL; an adult with income at or below 100% FPL). To be eligible for FPOS, individuals must not be covered by Medicaid for the Elderly, Blind, or Disabled or BadgerCare Plus and must be at or below 300% FPL. In December 2014, 65% of Wisconsin Medicaid members were enrolled in a health maintenance organization (HMO). A total of 118,309 eligible women who participated in one of 17 HMOs were included in the analysis. The results showed that 40% of clients ages 15-44 were provided most or moderately effective method of contraception; there was variation by HMO (34 to 50%). For more details, see the attached testing report.

Planned Parenthood of America (PPFA) is comprised of 66 independently incorporated affiliates, operating approximately 700 health centers in the United States, providing reproductive health care to nearly 2.7 million patients. De-identified, encounter-level data are captured in a quality information warehouse for a subset of affiliates. Data included in this analysis covers services provided at 25 affiliates between January 1 and December 31 2014. The final dataset analyzed included 838,872 female patients aged 15-44 years, who were cared for in one of 363 health centers, in the calendar year 2014. The results showed that 68% of clients ages 15-44 were provided a most or moderately effective method of contraception; there was variation by affiliate (28 to 90%) and health center. For more details, see the attached testing report.

The Title X Family Planning program was enacted in 1970, and is the only federal grant program dedicated solely to providing low-income individuals with comprehensive family planning and related preventive health services. The U.S. Department of Health and Human Services' Office of Population Affairs (OPA) oversees the Title X program. In 2014, grantees oversaw 4,100 family planning centers which served 4.2 million clients. Services are provided through state, county, and local health departments; community health centers; dedicated reproductive health centers; and hospital-based, school-based, faith-based, other private nonprofits. The scores were based on data reported to OPA in the annual Title X Family Planning Annual Report (FPAR), rather than claims data. The Title X data is included in this application to demonstrate that even in a program that is committed to the provision of family planning services, there is substantial room for improvement in the delivery of contraceptive services. The FPAR data has several advantages over claims data, in that it documents sterilization or LARC insertion in a year preceding the measurement year, and whether the client was seeking pregnancy. The results showed that 77 percent of clients ages 15-19 and 74% of clients ages 20-44 were provided a most or moderately effective method of contraception; there was variation by grantee (e.g., from 33 to 98% for adult clients). For more details, see the attached appendix.

It is desirable to have a very high percentage of women at risk of unintended pregnancy using a most or moderately effective method of contraception (e.g., more than 85%). Given the substantial room for improvement in the performance measure as described above, the focus will be on documenting improvement in the medium-term (e.g., a 15-percentage point increase over the next 4-5 years).

Due to some specific limitations of claims data, programs that serve a general population (such as a state Medicaid program) may need to take some additional steps if interested in approximating the actual percentage of women who are at risk of unintended pregnancy. The reason is because claims data do not capture the following aspects needed to determine if a woman is at risk of

unintended pregnancy: sexual experience, pregnancy intention, sterilization or LARC insertion in a year preceding the measurement year, and infecundity for non-contraceptive reasons (unless the woman had a procedure during the measurement year). These limitations can be partially addressed by using data from the National Survey of Family Growth (NSFG) which obtains detailed reproductive health data from women and men of reproductive age (see www.cdc.gov/nchs/nsfg). We conducted a special analysis of NSFG data so that the NSFG denominator matched the denominator identified with claims data (i.e., it excluded women who gave birth in the 2-month period prior to their interview, or who were still pregnant at the time of their interview). The NSFG estimates are found in Appendix B, with the first table comprised of all women of reproductive age, and the second table comprised of women who report Medicaid as their form of health care. By constructing a denominator for the NSFG data that matches the denominator obtained by claims data, we can approximate the percentages of women who are not at risk of unintended pregnancy, then use this information to adjust the claims based rates.

For example, the measures scores for Iowa Medicaid show that 28.0% of women ages 21-44 enrolled in the general Medicaid program were using a most or moderately effective method of contraception. NSFG estimates suggest that about 44% of adult women enrolled in Medicaid are not in need of contraceptive services because they have never had sex (1.8%), are seeking pregnancy (3.5%), are infecund for non-contraceptive reasons (4.7%), received LARC in a year preceding the measurement year (6.2%), or have been sterilized for contraceptive reasons in a year preceding the measurement year (28.1%). To adjust for the limitations of claims data, Iowa Medicaid might sum the measure score (28%) with the NSFG estimate of adult women not in need of contraceptive services (44%). This gives an adjusted estimate of 72% of adults whose contraceptive needs are met, and leaves up to a 28 percentage point opportunity for improvement.

The measure scores from programs that are focused on the delivery of reproductive health services (i.e., Iowa's and Wisconsin's family planning waiver programs, Planned Parenthood, and Title X) do not need to be adjusted with NSFG data. This is because the vast majority of clients who receive services from these programs are seeking contraceptive services and should therefore be considered at risk of unintended pregnancy.

1b.3. If no or limited performance data on the measure as specified is reported in 1b2, then provide a summary of data from the literature that indicates opportunity for improvement or overall less than optimal performance on the specific focus of measurement.

A special analysis of data from the National Survey of Family Growth (NSFG), 2011-2013, was conducted to examine contraceptive use patterns among women who were at risk of unintended pregnancy because they had ever had sex, were fecund, and were neither pregnant nor seeking pregnancy. The analysis showed that overall, only 43.5% of adolescents and 63% of adult women used a most or moderately effective method (CDC/NCHS, unpublished data).

1b.4. Provide disparities data from the measure as specified (current and over time) by population group, e.g., by race/ethnicity, gender, age, insurance status, socioeconomic status, and/or disability. *(This is required for endorsement maintenance. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities include.) This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*

1b.5. If no or limited data on disparities from the measure as specified is reported in 1b4, then provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement. Include citations.

We also conducted special analyses of the National Survey of Family Growth, 2011-2013, to explore disparities in the use of most and moderately effective methods of contraception. This analysis suggests that there are statistically significant differences by age and for women who were never married. However, there were no significant differences by race/ethnicity, most categories of marital status, and poverty level (CDC/NCHS, unpublished data). For more details, please see the testing report.

1c. High Priority (previously referred to as High Impact)

The measure addresses:

- a specific national health goal/priority identified by DHHS or the National Priorities Partnership convened by NQF; OR
- a demonstrated high-priority (high-impact) aspect of healthcare (e.g., affects large numbers of patients and/or has a substantial impact for a smaller population; leading cause of morbidity/mortality; high resource use (current and/or future); severity of illness; and severity of patient/societal consequences of poor quality).

1c.1. Demonstrated high priority aspect of healthcare

Affects large numbers, A leading cause of morbidity/mortality, Frequently performed procedure, High resource use, Patient/societal consequences of poor quality

1c.2. **If Other:**

1c.3. **Provide epidemiologic or resource use data that demonstrates the measure addresses a high priority aspect of healthcare. List citations in 1c.4.**

Over the course of a lifetime, most individuals will make decisions related to childbearing, i.e., how to prevent or achieve pregnancy so that they can attain their desired number and spacing of children. Of the 310 million people in the United States, 62 million (20%) are women of reproductive age, 15-44 years; approximately 20% are men of that same age (US Census Bureau).

Thirty-eight million women are in need of contraceptive services because they are at risk for unintended pregnancy – that is, they are sexually active, are able to get pregnant and want to avoid or space pregnancy (Frost 2015). Contraception is highly effective at preventing unintended pregnancy, and the most effective methods of long-acting reversible contraception (i.e., intrauterine devices and implants), have a failure rate that is less than 1% (Trussell 2011). Numerous studies have documented the cost-effectiveness of contraception, with $4-8 saved for every $1 invested (Cleland 2011, Frost 2008, Thomas 2012, Trussell 2012, Trussell 2013). Yet in 2011-2013, only 43.5% of adolescent and 63% of adult women at risk of unintended pregnancy were using a most or moderately effective method of contraception; and among those who use contraception, many do so inconsistently (CDC/NCHS special analysis, Trussell 2011, Jones 2012).

Due to these patterns of contraceptive use, the rate of unintended pregnancy is high in the United States. More than one-half (51%) of the 6.7 million pregnancies each year (3.2 million) are unintended (Finer & Zolna 2014). Despite recent reductions, each year more than 58 of every 1000 women aged 15-19 years become pregnant (Curtin 2015), and 250,000 adolescents give birth (Hamilton 2015). As a result, many teen mothers will achieve less education and lower incomes while their children may experience higher rates of negative outcomes such as poorer health, lowered academic achievement, and higher rates of teen pregnancy for female children and incarceration for male children (Hoffman & Maynard 2008, Sawhill et al 2014). Taxpayers also pay a high price for the nation's high rate of teen and unintended pregnancy. For example, the cost of teen pregnancy alone has been estimated at $9.4 billion per year (Hoffman 2010). Two-thirds of births resulting from unintended pregnancies among women of all ages—more than one million births—are publicly funded; the direct medical cost of those births was estimated at $21 billion in 2010 (this figure includes costs for prenatal care, labor and delivery, post-partum care, and one year of infant care) (Sonfield 2015). Further, family planning services can improve infant health, including a reduction in the rate of preterm and low birth weight infants (Tsui 2010, Gipson 2008, Conde-Agudela 2006, Shah 2011, Zhu 1999). Approximately 1 out of every 8 pregnancies in the United States results in preterm birth, and infant mortality rates remain high relative to other developed countries (MacDorman 2008, Martin 2015). Thus, by using contraceptive services to space births, and by offering preconception health services as part of family planning, it is possible to improve the health of infants, of women, and of men (CDC 2006).

The prevention of teen and unintended pregnancy and improved rates of birth spacing have repeatedly been identified as national priorities. Most recently, in 2015 the Institute of Medicine (IOM) recognized the importance of unintended pregnancy when they included it as one of 15 core measures that constitute the most vital signs for the nation's health and health care (IOM 2015). The United States' National Prevention Strategy (National Prevention Council) and Healthy People 2020 Objectives also include several specific objectives focused on unintended pregnancy and use of contraception (see www.healthypeople.gov/2020/topics-objectives/topic/family-planning). In 2015, a new Healthy People 2020 objective was approved that reports on the percentage of women at risk of unintended pregnancy using a most or moderately effective method of contraception, which is aligned with the proposed clinical performance measures.

1c.4. **Citations for data demonstrating high priority provided in 1a.3**

- American College of Obstetricians and Gynecologists, Committee on Gynecologic Practice. Increasing access to contraceptive implants and intrauterine devices to reduce unintended pregnancy. Committee Opinion Number 642; October 2015.
- American College of Obstetricians and Gynecologists (2011). Increasing Use of Contraceptive Implants and Intrauterine Devices to Reduce Unintended Pregnancy. ACOG Committee Opinion, Number 450, December 2009 (Reaffirmed 2011).
- American College of Obstetricians and Gynecologists (2012). Adolescents and Long-Acting Reversible Contraception: Implants and Intrauterine Devices. ACOG Committee Opinion, Number 539.
- American Academy of Pediatrics (2014). Contraception for Adolescents. Pediatrics, 134:e1244–e1256.
- CDC. Recommendations to improve preconception health and health care – United States. MMWR 2006;55(RR-06):1-23.
- CDC/OPA (2014). Providing Quality Family Planning Services: Recommendations of CDC and the US Office of Population Affairs, MMWR Recommendations and Reports, April 24, 2014, 63(RR04);1-29. Available online at:

http://www.cdc.gov/reproductivehealth/UnintendedPregnancy/QFP.htm.
• Cleland K, Peipert JF, Westhoff C, Spear S, Trussell J. Family planning as a cost-saving preventive health service. N Engl J Med 2011;364(18):e37.
• Conde-Agudelo A, Rosas-Bermudez A, Kafury-Goeta AC. Birth spacing and risk of adverse perinatal outcomes: a meta-analysis. Jama. Apr 19 2006;295(15):1809-1823.
• Curtin SC, Abma JC, Kost K. (2015). 2010 pregnancy rates among U.S. women. NCHS health e-stat.
• Finer LB, Zolna MR. Shifts in intended and unintended pregnancies in the United States, 2001-2008. Am J Public Health. 2014 Feb;
• Frost JJ et.al, Contraceptive Needs and Services, 2013 Update, New York: Guttmacher Institute, 2015, Available online at: www.guttmacher.org/pubs/win/.
• Frost J, Finer L, Tapales A. The Impact of Publicly Funded Family Planning Clinic Services on Unintended Pregnancies and Government Cost Savings. Journal of Health Care for the Poor and Underserved 2008;19(3):778-796.
• Gipson JD, Koenig MA, Hindin MJ (2008). The effects of unintended pregnancy on infant, child, and parental health: a review of the literature. Stud Fam Plann. 2008 Mar;39(1):18-38. Review.
• Hamilton BE, Martin JA, Osterman MJK, et al. Births: Final data for 2014. National vital statistics reports; vol 64 no 12. Hyattsville, MD: National Center for Health Statistics. 2015.
• Hoffman S, Maynard R. Kids having kids: Economic costs and social consequences of teen pregnancy 2nd ed. Washington, DC: Urban Institute Press 2008.
• Hoffman SD, 2010. Counting It Up: The Public Costs of Teen Childbearing. [cited 2014 February 7]; Available from: http://www.thenationalcampaign.org/costs/default.aspx
• Institute of Medicine, Committee on Core Metrics for Better Health at Lower Cost (2015). Vital Signs: Core Metrics for Health and Health Care Progress. The National Academy of Sciences, Washington DC.
• Jones J, Mosher W, Daniels K. Current contraceptive use in the United States, 2006–2010, and changes in patterns of use since 1995. National health statistics reports 2012(60).
• Macdorman M, Mathews T. Recent trends in infant mortality in the United States. In: NCHS data brief, no. 9. Hyattsville, MD: US Department of Health and Human Services, CDC; 2008.
• National Prevention Council. National Prevention Strategy. Washington DC: U.S. Department of Health and Human Services, Office of the Surgeon General, ; 2011.
• Sawhill I, Karpilow Q, Venator J (2014). The impact of unintended childbearing on future generations. Center on Children and Families at Brookings, Brookings Institution, Washington, DC.
• Shah PS, Balkhair T, Ohlsson A, Beyene J, Scott F, Frick C. Intention to become pregnant and low birth weight and preterm birth: a systematic review. Matern Child Health J 2011;15(2):205-16.
• Sonfield A and Kost K (2015). Public Costs from Unintended Pregnancies and the Role of Public Insurance Programs in Paying for Pregnancy and Infant Care: Estimates for 2010, the Guttmacher Institute. Available from http://www.guttmacher.org/pubs/public-costs-of-UP-2010.pdf.
• Thomas A. Three Strategies to Prevent Unintended Pregnancy. Journal of Policy Analysis and Management 2012;31(2):280-311.
• Trussell J. Contraceptive efficacy. In: Hatcher RA, Trussell J, Nelson AL, Cates W, Kowal D, editors. Contraceptive technology: twentieth revised edition. New York: Ardent Media; 2011, pp. 777–861.
• Trussell J. Contraceptive failure in the United States. Contraception 2011;83(5):397-404.
• Trussell J. Update on and correction to the cost-effectiveness of contraceptives in the United States. Contraception 2012;85(6):611.
• Trussell J, Henry N, Hassan F, Prezioso A, Law A, Filonenko A. Burden of unintended pregnancy in the United States: potential savings with increased use of long-acting reversible contraception. Contraception 2013;87(2):154-61.
• Tsui AO, McDonald-Mosley R, Burke AE (2010). Family planning and the burden of unintended pregnancies. Epidemiol Rev. 2010;32:152-74. Review.
• US Census Bureau. Age and Sex Composition in the United States: 2011. 2013 [cited 2013 October 17]; Available from: http://www.census.gov/population/age/data/2011comp.html
• US Department of Health and Human Services. Healthy People 2020 Objectives: Family Planning. [cited 2013 October 17]; Available from: http://www.healthypeople.gov/2020/topicsobjectives2020/overview.aspx?topicid=13
• Zhu BP, Rolfs RT, Nangle BE, Horan JM. Effect of the interval between pregnancies on perinatal outcomes. The New England journal of medicine. Feb 25 1999;340(8):589-594.

1c.5. If a PRO-PM (e.g. HRQoL/functional status, symptom/burden, experience with care, health-related behaviors), provide evidence that the target population values the measured PRO and finds it meaningful. *(Describe how and from whom their input*

was obtained.)
Not applicable.

2. Reliability and Validity—Scientific Acceptability of Measure Properties

Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. ***Measures must be judged to meet the subcriteria for both reliability and validity to pass this criterion and be evaluated against the remaining criteria.***

2a.1. Specifications The measure is well defined and precisely specified so it can be implemented consistently within and across organizations and allows for comparability. eMeasures should be specified in the Health Quality Measures Format (HQMF) and the Quality Data Model (QDM).

De.5. **Subject/Topic Area** *(check all the areas that apply):*
Perinatal and Reproductive Health : Newborn, Perinatal and Reproductive Health : Perinatal, Prevention

De.6. **Cross Cutting Areas** *(check all the areas that apply):*
Prevention

S.1. **Measure-specific Web Page** *(Provide a URL link to a web page specific for this measure that contains current detailed specifications including code lists, risk model details, and supplemental materials. Do not enter a URL linking to a home page or to general information.)*

S.2a. **If this is an eMeasure**, HQMF specifications must be attached. Attach the zipped output from the eMeasure authoring tool (MAT) - if the MAT was not used, contact staff. (Use the specification fields in this online form for the plain-language description of the specifications)
This is not an eMeasure **Attachment:**

S.2b. **Data Dictionary, Code Table, or Value Sets** *(and risk model codes and coefficients when applicable) must be attached. (Excel or csv file in the suggested format preferred - if not, contact staff)*
Attachment **Attachment:** Codes_2014_and_2015_MOST_MOD.xlsx

S.3. **For endorsement maintenance**, please briefly describe any changes to the measure specifications since last endorsement date and explain the reasons.

S.4. **Numerator Statement** *(Brief, narrative description of the measure focus or what is being measured about the target population, i.e., cases from the target population with the target process, condition, event, or outcome)*
IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
Women aged 15-44 years of age at risk of unintended pregnancy who are provided a most (sterilization, intrauterine device, implant) or moderately (pill, patch, ring, injectable, diaphragm) effective method of contraception.

S.5. **Time Period for Data** *(What is the time period in which data will be aggregated for the measure, e.g., 12 mo, 3 years, look back to August for flu vaccination? Note if there are different time periods for the numerator and denominator.)*
A twelve-month period of time is used.

S.6. **Numerator Details** *(All information required to identify and calculate the cases from the target population with the target process, condition, event, or outcome such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
IF an OUTCOME MEASURE, describe how the observed outcome is identified/counted. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
The target population is eligible women 15-44 years of age who are provided a most or moderately effective method of contraception. To identify the numerator, follow these steps:

Step 1 Define the numerator by identifying women who used a most (sterilization, IUD, implant) or moderately (injection, oral pills, patch, ring, or diaphragm) effective method of contraception in the measurement year. To do this, use the codes in Table UCM-E.

Step 2 Adjust for LARC removals and re-insertions. The LARC methods can be removed at the woman's request so adjustments must be made to reflect this. Use the codes in Table UCM-F to identify women who had their IUD or implant removed at any point during the measurement year. Check to see if they had an IUD or implant reinserted on the same or a subsequent date. If there is no code indicating reinsertion, use the codes in Table UCM-E to determine whether a woman was provided another most or moderately effective method. Do so by looking back over the 30 days prior to the removal (since a woman may receive a prescription for another method prior to the removal) as well as the period after the LARC removal (i.e., through the end of the measurement year). If there is no code for reinsertion or provision of another most or moderately effective method, consider them as a non-user.

Step 3 Calculate the rates by dividing the number of women who used a most or moderately effective method of contraception by the number of women in the denominator. Calculate the rates separately for adolescents and adults.

S.7. Denominator Statement *(Brief, narrative description of the target population being measured)*
Women aged 15-44 years of age who are at risk of unintended pregnancy.

S.8. Target Population Category *(Check all the populations for which the measure is specified and tested if any):*
Children's Health, Maternal Health

S.9. Denominator Details *(All information required to identify and calculate the target population/denominator such as definitions, specific data collection items/responses , code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
The target population is women of reproductive age (i.e., ages 15–44 years). In a Medicaid population, this includes:

- Women in the general Medicaid program who were continuously enrolled during the measurement year, i.e., had no more than one gap in enrollment of up to 45 days. To determine continuous enrollment for a Medicaid enrollee for whom enrollment is verified monthly, the enrollee may not have more than a 1-month gap in coverage (i.e., an enrollee whose coverage lapses for 2 months is not considered continuously enrolled)
- All women participating in a state-sponsored family planning-specific Section 1115 waiver or in a family–planning specific state plan amendment (SPA) program, even if they were not continuously enrolled. This is because the primary intent of these waiver and/or SPA programs is to provide family planning services, including contraception.

S.10. Denominator Exclusions *(Brief narrative description of exclusions from the target population)*
The following categories of women are excluded from the denominator: (1) those who are infecund for non-contraceptive reasons; (2) those who had a live birth in the last 2 months of the measurement year; or (3) those who were still pregnant or their pregnancy outcome was unknown at the end of the year.

S.11. Denominator Exclusion Details *(All information required to identify and calculate exclusions from the denominator such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
Follow the steps below to identify the denominator. The tables that are referenced are found in the attached Excel files (one file is for 2014 and the second is for 2015).

Step 1 Identify and exclude women who were infecund due to non-contraceptive reasons such as natural menopause or oophorectomy. To do this, use the codes listed in Table UCM-A.

Step 2 Identify women who were pregnant at any point in the measurement year by using the codes listed in Table UCM-B. We obtained this list of codes by reviewing the following documents, and including all pregnancy-related codes:

- CMS & NCHS (2011). ICD-9-CM Official Guidelines for Coding and Reporting, effective October 1, 2011. Available online at: http://www.cdc.gov/nchs/icd/icd9cm_addenda_guidelines.htm.
- CMS & NCHS (2016). ICD-10-CM Official Guidelines for Coding and Reporting FY 2016 Available online at: http://www.cdc.gov/nchs/icd/icd10cm.htm.

Step 3 Among women who were pregnant at any point in the measurement year, exclude those who:

- Had a live birth in the last 2 months of the measurement year because there may not have been an opportunity to provide

them with contraception. A two-month period was selected because the American College of Obstetricians and Gynecologists (ACOG) recommends having a postpartum visit by 6 weeks, and an additional 2 weeks was added to allow for reasonable delays in attending the postpartum visit. To identify live births, use the codes listed in Table UCM-D. This list of codes is drawn from the HEDIS measure of Prenatal and Postnatal care.

- Were still pregnant at the end of the year because they did not have a pregnancy outcome code indicating a non-live birth (Table UCM-C) or a live birth (Table UCM-D). Codes for non-live births were also drawn from the HEDIS measure of Prenatal and Postnatal Care.

Once the exclusions are applied, the denominator includes women who:

- Were not pregnant at any point in the measurement year,
- Were pregnant during the measurement year but whose pregnancy ended in the first 10 months of the measurement year, since there was adequate time to provide contraception in the postpartum period.
- Were pregnant during the measurement year but whose pregnancy ended in an ectopic pregnancy, stillbirth, miscarriage, or induced abortion.

S.12. **Stratification Details/Variables** *(All information required to stratify the measure results including the stratification variables, definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format with at S.2b)*

The primary stratification variable is age, so that adolescents can be examined separately from adult women. We recommend this for purposes of QI, rather than for purposes of risk stratification. Teen pregnancy is worthy of a separate focus because of the largepotential negative impact on the life of the teen and her child(ren), and the existence of unique programs and contraceptive counseling approaches tailored to this population. In the pilot data presented, we used age groups that are consistent with Center for Medicaid and CHIP Servies (CMCS) reporting requirements, i.e., adolescents are defined as 15-20 years and adults are 21-44 years of age.

S.13. **Risk Adjustment Type** (Select type. Provide specifications for risk stratification in S.12 and for statistical model in S.14-15)

No risk adjustment or risk stratification

If other:

S.14. **Identify the statistical risk model method and variables** *(Name the statistical method - e.g., logistic regression and list all the risk factor variables. Note - risk model development and testing should be addressed with measure testing under Scientific Acceptability)*

We do not believe that risk adjustment is justified. Although there are some variations in contraceptive use by socio-demographic characteristics, the reason for those patterns is based on modifiable clinical and programmatic considerations rather than differing biological responses to contraception. More detailed information about variations in use of most and moderately effective contraceptive methods by socio-demographic characteristics of women of reproductive age, can be found in the testing form (attached).

S.15. **Detailed risk model specifications** *(must be in attached data dictionary/code list Excel or csv file. Also indicate if available at measure-specific URL identified in S.1.)*

Note: Risk model details (including coefficients, equations, codes with descriptors, definitions), should be provided on a separate worksheet in the suggested format in the Excel or csv file with data dictionary/code lists at S.2b.

S.15a. **Detailed risk model specifications** (*if not provided in excel or csv file at S.2b*)

S.16. **Type of score:**

Rate/proportion

If other:

S.17. **Interpretation of Score** *(Classifies interpretation of score according to whether better quality is associated with a higher score, a lower score, a score falling within a defined interval, or a passing score)*

Better quality = Higher score

S.18. **Calculation Algorithm/Measure Logic** *(Describe the calculation of the measure score as an ordered sequence of steps including identifying the target population; exclusions; cases meeting the target process, condition, event, or outcome; aggregating data; risk adjustment; etc.)*

Step 1 Identify all women aged 15-44 years of age who were enrolled in the health plan or program. In the case of general Medicaid, include women who were continuously enrolled (i.e., had no more than one gap in enrollment of up to 45 days). In the case of women enrolled in a family planning-specific expansion program (1115 waiver or state plan amendment), include all women even if they do not meet the continuous enrollment criteria because the reason for their visit is related to pregnancy prevention.

Step 2 Define the denominator by excluding women who: (a) are infecund for non-contraceptive reasons; (b) had a live birth in the last 2 months of the measurement year; or (c) were still pregnant or their pregnancy outcome was unknown at the end of the year. Once exclusions are applied, the following groups of women will be included in the denominator: (a) those who were were not pregnant at any point in the measurement year; (b) those who had a live birth in the first 10 months of the measurement year; and (c) those who had a known miscarriage, stillbirth, ectopic pregnancy, or induced abortion during the measurement year.

Step 3 Define the numerator by using claims codes to identify women who adopted or continued use of one of the following methods of contraception in the measurement year: sterilization, IUD, implant, contraceptive injection, contraceptive pills, patch, ring, or diaphragm. Adjust for LARC removals, in the manner specified above.

Step 4 Calculate the rates by dividing the number who used a most or moderately effective method of contraception by the number of women in the denominator. Calculate the rates separately for adolescents and adults.

S.19. **Calculation Algorithm/Measure Logic Diagram URL or Attachment** *(You also may provide a diagram of the Calculation Algorithm/Measure Logic described above at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
Available in attached appendix at A.1

S.20. **Sampling** *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*
IF a PRO-PM, identify whether (and how) proxy responses are allowed.
The measure is based on data about all clients seen, not a sample.

S.21. **Survey/Patient-reported data** *(If measure is based on a survey, provide instructions for conducting the survey and guidance on minimum response rate.)*
IF a PRO-PM, specify calculation of response rates to be reported with performance measure results.
Not applicable

S.22. **Missing data** (specify how missing data are handled, e.g., imputation, delete case.)
Required for Composites and PRO-PMs.
Not applicable.

S.23. **Data Source** *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*
If other, please describe in S.24.
Administrative claims

S.24. **Data Source or Collection Instrument** *(Identify the specific data source/data collection instrument e.g. name of database, clinical registry, collection instrument, etc.)*
IF a PRO-PM, identify the specific PROM(s); and standard methods, modes, and languages of administration.
Adminisrative claims data are used to calculate the measure. The data request should include an eligibility file, paid and denied claims with diagnosis codes and procedures codes (HCPCS, CPT, and ICD-9-PCS/ICD-10-PCS), as well as NDC codes.

S.25. **Data Source or Collection Instrument** *(available at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
Available in attached appendix at A.1

S.26. **Level of Analysis** *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*

Facility, Health Plan, Population : Regional, Population : State

S.27. **Care Setting** *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*

If other:

S.28. **COMPOSITE Performance Measure** - Additional Specifications *(Use this section as needed for aggregation and weighting rules, or calculation of individual performance measures if not individually endorsed.)*

2a. Reliability – See attached Measure Testing Submission Form
2b. Validity – See attached Measure Testing Submission Form
nqf_testing_Contraceptive_Care_MOST_MOD_v3-635918257250215078.docx

Measure Number (*if previously endorsed*)**:** Click here to enter NQF number

Measure Title: Contraceptive Care – Most and moderately effective methods

Date of Submission: Click here to enter a date

Type of Measure: Intermediate Clinical Outcome

☐ Composite – ***STOP – use composite testing form***	☒ Outcome (*including PRO-PM*)
☐ Cost/resource	☐ Process
☐ Efficiency	☐ Structure

Instructions

- Measures must be tested for all the data sources and levels of analyses that are specified. ***If there is more than one set of data specifications or more than one level of analysis, contact NQF staff*** about how to present all the testing information in one form.
- **For all measures, sections 1, 2a2, 2b2, 2b3, and 2b5 must be completed.**
- **For outcome and resource use measures**, section **2b4** also must be completed.
- If specified for **multiple data sources/sets of specificaitons** (e.g., claims and EHRs), section **2b6** also must be completed.
- Respond to all questions as instructed with answers immediately following the question. All information on testing to demonstrate meeting the subcriteria for reliability (2a2) and validity (2b2-2b6) must be in this form. An appendix for *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 20 pages (*incuding questions/instructions;* minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.
- For information on the most updated guidance on how to address sociodemographic variables and testing in this form refer to the release notes for version 6.6 of the Measure Testing Attachment.

Note: The information provided in this form is intended to aid the Steering Committee and other stakeholders in understanding to what degree the testing results for this measure meet NQF's evaluation criteria for testing.

2a2. Reliability testing [10] demonstrates the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise. For **PRO-PMs and composite performance measures**, reliability should be demonstrated for the computed performance score.

2b2. Validity testing [11] demonstrates that the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality. For **PRO-PMs and composite performance measures**, validity should be demonstrated for the computed performance score.

2b3. Exclusions are supported by the clinical evidence; otherwise, they are supported by evidence of sufficient frequency of occurrence so that results are distorted without the exclusion; [12]

AND

If patient preference (e.g., informed decision-making) is a basis for exclusion, there must be evidence that the exclusion impacts performance on the measure; in such cases, the measure must be specified so that the information about patient preference and the effect on the measure is transparent (e.g., numerator category computed separately, denominator exclusion category computed separately). [13]

2b4. For outcome measures and other measures when indicated (e.g., resource use):

- **an evidence-based risk-adjustment strategy** (e.g., risk models, risk stratification) is specified; is based on patient factors (including clinical and sociodemographic factors) that influence the measured outcome and are present at start of care; [14,15] and has demonstrated adequate discrimination and calibration

OR

- rationale/data support no risk adjustment/ stratification.

2b5. Data analysis of computed measure scores demonstrates that methods for scoring and analysis of the specified measure allow for **identification of statistically significant and practically/clinically meaningful** [16] **differences in performance**;

OR

there is evidence of overall less-than-optimal performance.

2b6. If multiple data sources/methods are specified, there is demonstration they produce comparable results.

2b7. For **eMeasures, composites, and PRO-PMs** (or other measures susceptible to missing data), analyses identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias.

Notes

10. Reliability testing applies to both the data elements and computed measure score. Examples of reliability testing for data elements include, but are not limited to: inter-rater/abstractor or intra-rater/abstractor studies; internal consistency for multi-item scales; test-retest for survey items. Reliability testing of the measure score addresses precision of measurement (e.g., signal-to-noise).

11. Validity testing applies to both the data elements and computed measure score. Validity testing of data elements typically analyzes agreement with another authoritative source of the same information. Examples of validity testing of the measure score include, but are not limited to: testing hypotheses that the measures scores indicate quality of care, e.g., measure scores are different for groups known to have differences in quality assessed by another valid quality measure or method; correlation of measure scores with another valid indicator of quality for the specific topic; or relationship to conceptually related measures (e.g., scores on process measures to scores on outcome measures). Face validity of the measure score as a quality indicator may be adequate if accomplished through a systematic and transparent process, by identified experts, and explicitly addresses whether performance scores resulting from the measure as specified can be used to distinguish good from poor quality.

12. Examples of evidence that an exclusion distorts measure results include, but are not limited to: frequency of occurrence, variability of exclusions across providers, and sensitivity analyses with and without the exclusion.

13. Patient preference is not a clinical exception to eligibility and can be influenced by provider interventions.

14. Risk factors that influence outcomes should not be specified as exclusions

15. With large enough sample sizes, small differences that are statistically significant may or may not be practically or clinically meaningful. The substantive question may be, for example, whether a statistically significant difference of one percentage point in the percentage of patients who received smoking cessation counseling (e.g., 74 percent v. 75 percent) is clinically meaningful; or whether a statistically significant difference of $25 in cost for an episode of care (e.g., $5,000 v. $5,025) is practically meaningful. Measures with overall less-than-optimal performance may not demonstrate much variability across providers.

1. DATA/SAMPLE USED FOR ALL TESTING OF THIS MEASURE

Often the same data are used for all aspects of measure testing. In an effort to eliminate duplication, the first five questions apply to all measure testing. If there are differences by aspect of testing,(e.g., reliability vs. validity) be sure to indicate the specific differences in question 1.7.

1.1. What type of data was used for testing? (*Check all the sources of data identified in the measure specifications and data used for testing the measure. Testing must be provided for all the sources of data specified and intended for measure implementation.* ***If different data sources are used for the numerator and denominator, indicate N [numerator] or D [denominator] after the checkbox.***)

Measure Specified to Use Data From: (*must be consistent with data sources entered in S.23*)	**Measure Tested with Data From:**
☐ abstracted from paper record	☐ abstracted from paper record
☒ administrative claims	☒ administrative claims
☐ clinical database/registry	☐ clinical database/registry
☐ abstracted from electronic health record	☐ abstracted from electronic health record
☐ eMeasure (HQMF) implemented in EHRs	☐ eMeasure (HQMF) implemented in EHRs
☐ other: Click here to describe	☐ other: Click here to describe

1.2. If an existing dataset was used, identify the specific dataset (*the dataset used for testing must be consistent with the measure specifications for target population and healthcare entities being measured; e.g., Medicare Part A claims, Medicaid claims, other commercial insurance, nursing home MDS, home health OASIS, clinical registry*).

Claims data from three programs were used for testing:

(1) The Planned Parenthood Federation of America (PPFA). In 2014, PPFA comprised 66 independently incorporated affiliates, operating approximately 700 health centers in the United States, and providing reproductive health care to nearly 2.7 million patients. De-identified, encounter-level data are captured in a quality information warehouse for a subset of affiliates. The final dataset analyzed included 838,872 female patients aged 15-44 years, who received services from 25 PPFA affiliates between January 1 and December 31, 2014. The measures were evaluated using all claims data among the eligible population, which included de-identified patient encounters, and identifiers for billing providers and health centers within affiliates. Affiliates cover geographic service areas that range from several counties within a state, a state population, and multiple states. Among the 25 affiliates included in our dataset, there were 363 health centers, and 4,467 unique billing providers nested among the health centers. These data cover diverse geographic regions and extremely large member populations, and thus may be considered reasonably representative of the U.S. population of women of reproductive age. Hence, OPA suggests that the affiliate be considered a reasonable proxy for a U.S. state, for purposes of this application.

(2) The Iowa Medicaid Program (IME). The IME dataset comprised all female Medicaid clients aged 15-44 years who resided in 6 public health regions, participated in either fee-for-service care or in two health plans, and participated in either the general Medicaid program or the state's family planning waiver program. Iowa's Medicaid Enterprise (IME) provides contraceptive services to women through its general Medicaid program and its family planning waiver program (IFPN). Services are available to Iowa residents who are US citizens or qualified immigrants. To be eligible for IFPN services, the following guidelines apply: an individual does not have insurance or your insurance does not cover family planning services; the individual is a man or woman between the ages of 12 and 54; family income is at or below 300 percent of the federal poverty level; and women whose pregnancy and delivery was covered by Medicaid will have family planning services covered. In 2013, Medicaid services in Iowa were provided primarily on a fee-for-service basis, although a small percentage of clients (approximately 2%) were provided care through one of two managed care organizations (MCO).

(3) The Wisconsin Medicaid Program (WMP). The WMP dataset is comprised of all female Medicaid clients aged 15-44 years who in 2014 resided in Wisconsin, had a paid Medicaid claim, and participated in either the general Medicaid program or the state's Family Planning Only Services (FPOS) program. The WMP provides contraceptive services to women through its general Medicaid program (BadgerCare Plus) and FPOS. FPOS members receive services on a fee for service basis only. Services are available to Wisconsin residents who are US citizens or qualified immigrants meeting income eligibility criteria (e.g., a child <18 years with household income at or below 300% FPL; an adult with income at or below 100% FPL). To be eligible for FPOS, individuals must not be covered by Medicaid for the Elderly, Blind, or Disabled or BadgerCare Plus and must be at or below 300% FPL. In December 2014, 65% of Wisconsin Medicaid members were enrolled in one of 18 health maintenance organizations (HMO).

1.3. What are the dates of the data used in testing? January 1 2013 – December 31 2014

Data from PPFA covered the period January 1 2014 – December 31 2014.
Data from IME covered the period January 1, 2013 – December 31, 2013.
Data from Wisconsin Medicaid covered the period January 1, 2014 – December 31, 2014.

1.4. What levels of analysis were tested? (*testing must be provided for all the levels specified and intended for measure implementation, e.g., individual clinician, hospital, health plan*)

Measure Specified to Measure Performance of: (*must be consistent with levels entered in item S.26*)	**Measure Tested at Level of:**
☐ individual clinician	☐ individual clinician
☐ group/practice	☐ group/practice
☒ hospital/facility/agency	☒ hospital/facility/agency
☒ health plan	☒ health plan
☒ other: Population/state equivalent, public health region, benefit type	☒ other: Population/state equivalent, public health region, benefit type

1.5. How many and which measured entities were included in the testing and analysis (by level of analysis and data source)? (*identify the number and descriptive characteristics of measured entities included in the analysis (e.g., size, location, type); if a sample was used, describe how entities were selected for inclusion in the sample*)

Reliability
The measure was tested at several levels, as shown in the table below.

Level	Number of measured entities	Data Source
Affiliate	25	PPFA
Health center	363	PPFA
Benefit type (general Medicaid vs FP waiver)	2	IME
Public health region	6	IME
Health plan (Medicaid health maintenance organization)	17	WMP

Validity
A panel of experts assessed the measure's face validity.

1.6. How many and which patients were included in the testing and analysis (by level of analysis and data source)? (*identify the number and descriptive characteristics of patients included in the analysis (e.g., age, sex, race, diagnosis); if*

a sample was used, describe how patients were selected for inclusion in the sample)

Level of analysis	Number of patients		
	15 - 20 years	**21 -44 years**	**15 - 44 years**
Affiliate (PPFA), n=25			
TOTAL	203,970	634,902	838,872
Range	294 - 42,698	1265 – 131,187	1701 – 173,885
Health centers within affiliate (PPFA), n=363			
TOTAL	203,970	634,902	838,872
Range	8 - 2984	31 – 11,391	48 – 13,335
Type of benefit (IME)			
General Medicaid	5254	9483	14737
Family planning waiver	6445	23568	30013
TOTAL	**11,699**	**33,051**	**44,750**
Public health region (IME)			
Region 1	3460	9588	13048
Region 2	1154	2906	4060
Region 3	1176	3175	4351
Region 4	1087	2887	3974
Region 5	1701	4359	6060
Region 6	3121	10136	13257
TOTAL	**11,699**	**33,051**	**44,750**
Health plan (WMP)			
HMO 1	4832	14043	18875
HMO 2	1838	5688	7526
HMO 3	920	2862	3782
HMO 4	1795	5681	7476
HMO 5	1231	3936	5167
HMO 6	219	725	944
HMO 7	558	1608	2166
HMO 8	352	1096	1448
HMO 9	1623	6164	7787
HMO 10	618	1683	2301
HMO 11	4898	15166	20064
HMO 12	1239	4290	5529
HMO 13	269	853	1122
HMO 14	2149	5596	7745
HMO 15	56	240	296
HMO 16	5114	18875	23989
HMO 17	559	1533	2092
TOTAL	**28,270**	**90,039**	**118,309**

1.7. If there are differences in the data or sample used for different aspects of testing (e.g., reliability, validity, exclusions, risk adjustment), identify how the data or sample are different for each aspect of testing reported below.

Not applicable

1.8 What were the patient-level sociodemographic (SDS) variables that were available and analyzed in the data or sample used? For example, patient-reported data (e.g., income, education, language), proxy variables when SDS data are not collected from each patient (e.g. census tract), or patient community characteristics (e.g. percent vacant housing, crime rate).

We assessed reliability of the measures after stratifying by age, i.e., adolescent versus adult. Teen pregnancy is worthy of a separate focus because of the large potential negative impact on the life of the teen and her child(ren), and the existence of unique programs and contraceptive counseling approaches tailored to this population. To define age groups, we used the categories developed by the Center for Medicaid and CHIP Services (CMCS), i.e., individuals aged 15 through 20 years (15-20) were defined as adolescents, and individuals aged 21 through 44 years (21-44) were defined as adults.

_____

2a2. RELIABILITY TESTING

***Note**: If accuracy/correctness (validity) of data elements was empirically tested, separate reliability testing of data elements is not required – in 2a2.1 check critical data elements; in 2a2.2 enter "see section 2b2 for validity testing of data elements"; and skip 2a2.3 and 2a2.4.*

2a2.1. What level of reliability testing was conducted? (*may be one or both levels*)

☐ **Critical data elements used in the measure** (*e.g., inter-abstractor reliability; data element reliability must address ALL critical data elements*)

☒ **Performance measure score** (e.g., *signal-to-noise analysis*)

2a2.2. For each level checked above, describe the method of reliability testing and what it tests (*describe the steps—do not just name a method; what type of error does it test; what statistical analysis was used*)

Several methods have been suggested to assess the reliability of provider-level performance measures (Adams, 2010; Scholle et al, 2008; Fung et al, 2010). These methods may focus on different facets of reliability such as consistency across time, consistency across raters or units, or variability at different levels of aggregation. The NQF has suggested a ***signal-to-noise*** approach as one way to evaluate measure reliability. According to Adams (2009), reliability can be assessed by the proportion of variance in a performance measure due to systemic differences across measured units (signal) in relation to random error (noise) within units.

When analytic units fall into a natural hierarchy (e.g. clients nested within health centers nested within health plan organizations), one can estimate multilevel variance components using hierarchical generalized linear modeling (HGLM) (Raudenbush and Bryk, 2002; Woltman et al, 2012). In this approach the within-provider regression coefficients are allowed to vary across providers as random effects. The covariance parameter for the random effect estimates the true between-provider variance after accounting for within-provider variance. HGLM methods are robust and well-developed for continuous outcomes, and have more recently been applied to binary outcomes (Ridout, 1999; Molenberghs et al, 2007).

In the present analyses, multi-level mixed models were fit to each dataset using a hierarchical SAS 9.3 GLIMMIX procedure with a log link function. Parameters were estimated by pseudo-maximum-likelihood using the Laplace method (Ene et al, 2012). Modeling proceeded in a top-down manner starting from the largest unit of aggregation; the variance component (random coefficient) was always estimated for the top level.

Reliability was then calculated as a function of the intraclass correlation (ICC) and the median number cases per unit, using the Spearman-Brown prophecy (Eijkenaar et al, 2013). ICCs are derived using the estimated variance component for the level of interest divided by the total variance (Wu et al, 2012; He et al, 2014). ICCs conceptually represent the proportion of total variation accounted for by the between-provider level, and thus follows the signal-to-noise framework suggested by NQF.

The HGLM method of estimation assumes a normally distributed error component; some authors have noted that ICCs on the logit scale can be inflated under certain circumstances when population rates are near the extremes (Wu et al,

2012). To provide more conservative estimation, medians were used in the Spearman-Brown reliability formula; the use of means would tend to bias estimates upward due to one or two atypically large provider units.

Structure of the Data

PPFA dataset. PPFA affiliates offer services within health centers. Inside each health center a group of billing providers offer care to clients. Modeling began at the topmost affiliate level (n=25), where all clients were aggregated within affiliate for the calculation of rate of most/moderately effective contraceptive use. The next level of analysis was performed within each of the 25 affiliates to examine health center rates (n=363 across all affiliates). This provided a basic 2-level structure of clients aggregated within each hierarchical unit. The top-down modeling approach enabled us to ignore small sample size problems and attribution error among individual billing providers; it also allowed us to explore the lowest level of 'granularity' for distinguishing performance among health centers of smaller size.

Iowa Medicaid Enterprise dataset. For IME data, modeling similarly proceeded from the top down starting with public health region (n=6). Unlike the PPFA data, IME data could not be examined by health facility. Instead the analysts were interested in reporting on public health region and benefit type (family planning waiver or general Medicaid benefit). Since the benefit type categories exist across regions, there is no nesting of units. Therefore, in Iowa the six regions were simply crossed with the type of benefit (n=12). Both of these crossed analyses were thought to provide useful and potentially actionable information about the interplay of regional and administrative influences on service delivery.

Wisconsin Medicaid dataset. For WMP data, modeling similarly proceeded from the top down starting with health maintenance organization (data from 17 of 18 HMOs was available).

2a2.3. For each level of testing checked above, what were the statistical results from reliability testing? (e.*g., percent agreement and kappa for the critical data elements; distribution of reliability statistics from a signal-to-noise analysis*)

The table below shows summary results of the reliability analyses at five levels (i.e., affiliate, health center, health plan, public health region and region by benefit type), stratified by three age categories (i.e., 15-20, 21-44, and 15-44). More detailed information about the analyses at each level can be found in Tables 1-4 (appended at the end of the form).

Level	Age group	Results		
		Median N	ICC	Reliability
Affiliate (PPFA)	15-20	4839	.1164	.9984
	21-44	11648	.1232	.9994
	15-44	16590	.1191	.9996
Health centers (PPFA) (estimated within each affiliate)	15-20	366	.0612 (median)	.6096 - .9985
	21-44	1016	.0484 (median)	.6709 - .9990
	15-44	1379	.0581 (median)	.7056 - .9997
Public health region (IME)	15-20	1438	.0121	.9461
	21-44	3767	.0041	.9399
	15-44	5205	.0034	.9461
Benefit type (IME)	15-20	11699	.1268	.9988
	21-44	33051	.0057	.9895
	15-44	44750	.0463	.9991
Region by benefit type (IME)	15-20	716	.1929	.9942
	21-44	2325	.2148	.9984
	15-44	2954	.1920	.9986
Health plan/HMO (WMP)	15-20	1231	.0017	.6767
	21-44	3936	.0029	.9206

	15-44	5167	.0018	.9048

For each level, the overall reliability was estimated using the medians as previously mentioned. ICCs, an indicator of the proportion of variance explained by the groupings, are also shown. Similar studies of hierarchical binary outcomes estimate ICCs in a typical range of .02 - .18 (Fung et al, 2010). The moderate ICCs found in our analyses, combined with the large volume of patients at most levels, tend to generate high reliability estimates. Using the 'floor' of reliability, we also calculate the minimum number of cases required to achieve acceptable reliability thresholds for each level.

The estimated reliabilities remain above .90 for affiliates, for most affiliate groupings of health centers, for region, for benefit type, for region by benefit type, and 2 of 3 age groups at the health plan level. The ICCs at these levels were variable, ranging from low (e.g., <1%) to high (up to 21%). Of note, reliability did decline slightly in the analysis of the health centers within each affiliate and for the 15-20 age group at the health plan level. This would be expected since the volume of patients decreased, the cases per unit were less stable, and the rates were slightly more consistent among the health centers. The estimated reliabilities remain above .90 for most affiliates, and most levels below affiliate, due to sufficient patient volume. An exception occurred with two of the affiliates that contained only a single health center. Since there can be no variance in rates for a single unit, the health center level ICCs (and reliabilities) for those two affiliates are not included above.

It is commonly advised that reliability should be $\geq$ .90 for making decisions, and $\geq$ .70 for general reporting/monitoring (Eijkenaar, 2013; Adams, 2010). The Spearman-Brown prophecy allows one to test different values for ICC and patient volume per unit in order to predict expected reliability. Using an ICC value near the 20th percentile as a conservative expected correlation within units, we can compute the minimum recommended case load per level for each threshold of reliability. For example, for within-affiliate reporting of health centers, we used a conservative expected floor of .02 ICC to recommend that health centers have at least 115 patient cases for reporting rates to maintain >.70 reliability, and 450 cases to maintain >.90 reliability. The median ICC from actual data was nearly 3 times our conservative floor value (and most health centers exceeded this minimum number of cases) thus our reported reliabilities were considerably higher.

2a2.4 What is your interpretation of the results in terms of demonstrating reliability? (i.*e., what do the results mean and what are the norms for the test conducted?*)

Despite the challenges of recoding claims data to obtain contraceptive rates, having large and diverse datasets available made a positive impact on reliability. For the PPFA data both at the affiliate level and at the next level down (groups of health centers within affiliate), we found reliabilities well above the commonly accepted .90 reliability threshold for reporting and decision-making. This was largely driven by two factors. First, the data exhibited adequate variation in the rates of contraceptive use at both the affiliate and lower levels. Second, the number of patients per unit at the affiliate level was mostly in the thousands, and at the lower levels, usually exceeded several hundred. For the IME data, the rates were much more uniform by region resulting in lower ICCs, but the volume of clients still enabled adequate reliability for distinguishing performance. When region was crossed by type of health plan or benefit the contraceptive rates were more variable among the units, so even given the smaller size of these analytic units the estimated reliabilities were higher.

In performing this analysis, we attempted to provide a conservative estimate of reliability wherever possible. Using medians rather than means, and presenting the 'floor' of reliability that may be observed for the smallest units, we bracket the results with worst-case scenarios. We further utilized a conservative value of ICC to recommend minimum patients per unit to maintain the .70 and .90 levels of reliability. In future years, analyses could examine the actual ICCs in order to make appropriate determinations about cases per unit. Yet even with these conservative methods, the 2014 data at the affiliate (state) and lower levels appears to provide sufficient reliability for reporting contraceptive rates.

2b2. VALIDITY TESTING

2b2.1. What level of validity testing was conducted? (*may be one or both levels*)

☐ **Critical data elements** (*data element validity must address ALL critical data elements*)

☒ **Performance measure score**

☐ **Empirical validity testing**

☒ **Systematic assessment of face validity of performance measure score as an indicator** of quality or resource use (*i.e., is an accurate reflection of performance on quality or resource use and can distinguish good from poor performance*)

2b2.2. For each level of testing checked above, describe the method of validity testing and what it tests (*describe the steps—do not just name a method; what was tested, e.g., accuracy of data elements compared to authoritative source, relationship to another measure as expected; what statistical analysis was used)*

We used a systematic process to assess the face validity of the performance measure, i.e., whether the corresponding measure scores correctly reflect the quality of care provided and adequately identify differences in quality. Nine experts with the following characteristics were identified: (1) expertise in the delivery of contraceptive services, as evidenced by employment in a clinical or managerial capacity for at least 3 years during which they delivered contraceptive services in a clinical setting (i.e., public and private family planning and primary care providers, or health administrators); and (2) expertise in the use of performance measures, as evidenced by participation in at least one effort to collect and use performance measurement data for the purpose of improving clinical services in the setting(s) in which they work. Below is the final list of experts who participated in the assessment:

1. Carol Brady, MA, Project Director, Florida Association of Healthy Start Coalitions, Inc.
2. Anne Burke, MD, Associate Professor, School of Medicine, Johns Hopkins Bayview Medical Center
   Vanessa Dalton, MD, MPH, Associate Professor, Director, Program on Women's Health Care Effectiveness Research, University of Michigan
3. Anne Dunlop, MD, MPH, Program Director, Preventive Medicine Division, Emory University School of Medicine
4. Daryn Eikner, MS, Vice President of Health Care Delivery, National Family Planning & Reproductive Health Association
5. Jan Engstrom, PhD, RN, CNM, WHNP-BC, Professor & Acting Chairperson, Department of Women, Children and Family Nursing, College of Nursing, Armour Academic Center
6. Mark Hathaway, MD, MPH, Senior Technical Advisor, Jhpiego – Johns Hopkins University
7. Michael Policar, MD, MPH, Clinical Professor of Obstetrics, Gynecology, and Reproductive Sciences, UCSF School of Medicine
8. Linda Wheal, Maternal Health Program Manager, Bureau of Quality Management, Illinois Department of Healthcare and Family Services

We contacted the selected experts to confirm consent to participate via email. Each expert panelist was sent a disclosure form to report any relevant financial or other competing interests; disclosures were compiled with brief biographies and shared with all panelists. Upon receipt of the disclosure form we sent the participant information about the measure specifications and other background information about the measure. Participants then participated in a webinar designed to provide important background information about the measure, how it is computed, the NQF endorsement process, and how the face validity assessment will be used in the application package that will be submitted to NQF. After reviewing the measure specifications and participating in the webinar the participants completed a survey (anonymous) that asked the following question about the measure:

The scores obtained from the measure as specified will provide an accurate reflection of quality and can be used to distinguish good and poor quality in contraceptive services:

1= Strongly Disagree
3=Neither Agree nor Disagree
5= Strongly Agree

ICD-10 Conversion:

We tested the measure specifications based on 2014 codes, but have also included the codes needed to calculate the measure using ICD-10 and 2015 NDC codes. Both sets of codes are attached. Our goal was to convert the measure to a new code set, fully consistent with the intent of the original measure. A description of how we converted from ICD-9 to ICD-10 is provided below, for each table in the measure specifications.

- **Sterilization for non-contraceptive reasons (Table UCM-A)**

We identified the 2015 ICD-10 codes for this table by using ICD-10 online conversion tools and confirming codes in the ICD-10-CM Expert for Physicians complete official code set, as well as with a clinical expert. These were confirmed with a clinical expert, Denise Wheeler, MS, Family Planning Director at the Iowa Department of Public Health.

- **Pregnancy codes (Table UCM-B)**

We identified the 2015 ICD-10 codes for this table by searching the NCHS/CMS publication, "ICD-10-CM Official Guidelines for Coding and Reporting, FY 2015". Pregnancy-related codes were found in "Chapter 15: Pregnancy, Childbirth and the Puerperium (O00-O9A)", and also Z codes for "outcome of delivery".

- **Known miscarriage, ectopic pregnancy, stillbirth, or induced abortion (Table UCM-C)**

These codes were identified by copying the Non-live Births Value Set from NCQA's Prenatal & Postpartum Care (PPC) measure (NQF#1517), as well as non-live birth codes in "Chapter 15: Pregnancy, Childbirth and Purperium (O00-O9A)". In the PPC measure, these codes are used to identify live births.

- **Delivery resulting in a live birth (Table UCM -D)**

These codes were identified by copying the Deliveries Value Set from NCQA's Prenatal & Postpartum Care (PPC) measure (NQF#1517), excluding extraction of products of conception retained and ectopic. In the PPC measure, these codes are used to identify live births.

- **Contraceptive codes (Tables UCM E, F and G)**

We used ICD-10 online conversion tools and confirming codes in the ICD-10-CM Expert for Physicians complete official code set. They were cross-checked against a ICD-10 conversion chart for family planning services that was prepared by Dr Michael Policar, from the University of California-San Francisco, and confirmed with a clinical expert, Denise Wheeler, MS, Family Planning Director at the Iowa Department of Public Health. NDC codes for 2015 were updated by using the codes for contraception contained in the HEDIS specifications for Chlamydia screening.

2b2.3. What were the statistical results from validity testing? (*e.g., correlation; t-test*)

The mean rating from the face validity assessment for this measure was 4.67 with a median of 5 (Strongly Agree), range 4-5. There were 66.7% (n = 6) of respondents who strongly agreed and 33.3% (n = 3) of respondents who agreed that the scores obtained from this measure, as specified, will provide an accurate reflection of quality and can be used to distinguish good and poor quality in contraceptive services. One respondent replied that he or she thinks that "the proposed measures are valid measure of quality contraceptive care for healthy women" and one responded he or she "feels STRONGLY that the adoption of these measures will promote providers' and practices' attention to reproductive planning and contraceptive care as part and parcel of women's primary health care." One respondent strongly agrees "that the measure has excellent face validity as currently specified." He or she also responded, "However, in the future, we would suggest considering the use of a look-back period using claims data to identify previous use of long-acting contraceptives." One respondent pointed out that "quality of the indicator will in part depend on how well 'unintended' is characterized."

2b2.4. What is your interpretation of the results in terms of demonstrating validity? (i.*e., what do the results mean and what are the norms for the test conducted?*)

We think that the responses to the face validity assessment indicate that the measure will provide an accurate reflection of quality and can be used to distinguish good and poor quality in contraceptive services.

_____

2b3. EXCLUSIONS ANALYSIS

NA ☐ no exclusions — *skip to section 2b4*

2b3.1. Describe the method of testing exclusions and what it tests (*describe the steps—do not just name a method; what was tested, e.g., whether exclusions affect overall performance scores; what statistical analysis was used*)

Exclusions were not formally tested. The rationale for exclusion was due to the fact that some women are not at risk of unintended pregnancy due to infecundity or pregnancy.

2b3.2. What were the statistical results from testing exclusions? (*include overall number and percentage of individuals excluded, frequency distribution of exclusions across measured entities, and impact on performance measure scores*)

The table below shows the number of women excluded in each of the two datasets, presented by the reason for exclusion.

		Number of women		
		PPFA, 2014	IME, 2013	WMP, 2014
Women 15-44 years of age		950,647	49,232	132,940
Exclusions	Infecund for non-contraceptive reasons	83	169	2,025
	Had a live birth in the last 2 months of the measurement year	7	520	2,995
	Pregnant or their pregnancy outcome was unknown at the end of the measurement year	111,685	3793	9,611
Number of women 15-44 years of age, after exclusions		838,872	44,750	118,309

2b3.3. What is your interpretation of the results in terms of demonstrating that exclusions are needed to prevent unfair distortion of performance results? (*i.e., the value outweighs the burden of increased data collection and analysis. Note: **If patient preference is an exclusion**, the measure must be specified so that the effect on the performance score is transparent, e.g., scores with and without exclusion*)

When combined, the total number of exclusions in each of the two data sets comprised 11.8% (PPFA), 9.1% (IME) and 11% (WMP) of all women 15-44 years of age, although the relative contribution of each type of exclusion varied by data set (e.g.., live births in the last 2 months of the year were a relatively larger population in IME dataset than the PPFA dataset). These differences are likely explained by the fact that the emphasis of each program is slightly different, with the PPFA program more heavily focused on delivery of reproductive health care while the IME and WMP programs offer a wider range of primary, acute and curative care services. The number of women excluded will have a noticeable impact on the rates, and will be important to reassure providers that the measure is as 'fair' in terms of identifying the population at risk as claims data will allow it to be. For these reasons, we believe that the burden of applying the exclusion criteria is outweighed by the benefits of doing so.

_____

2b4. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES

If not an intermediate or health outcome, or PRO-PM, or resource use measure, skip to section 2b5.

2b4.1. What method of controlling for differences in case mix is used?

☒ **No risk adjustment or stratification**

☐ **Statistical risk model with** Click here to enter number of factors **risk factors**

☐ **Stratification by** Click here to enter number of categories **risk categories**

☐ **Other,** Click here to enter description

2b4.2. If an outcome or resource use measure is not risk adjusted or stratified, provide rationale and analyses to demonstrate that controlling for differences in patient characteristics (case mix) is not needed to achieve fair comparisons across measured entities.

We do not believe that risk adjustment is justified. Although there are potential variations in contraceptive use by socio-demographic characteristics, the reason for those patterns is based on modifiable clinical and programmatic considerations rather than differing biological responses to contraception. Although providers may see some local variations by socio-demographic characteristics, we do not believe that these differences will be maintained if contraceptive services are offered in a client-centered manner, as defined by CDC-OPA recommendations for providing quality family planning services (CDC-OPA, 2014).

A special analysis of data from the National Survey of Family Growth (NSFG), 2011-2013, was conducted to explore disparities in the use of most and moderately effective methods of contraception (see table below). This analysis suggests that there are statistically significant differences by age and for women who were never married. However, there were no significant differences by race/ethnicity, most categories of marital status, and poverty level.

Percentage of women 15-44 years of age at risk of unintended pregnancy*
that used a most or moderately effective method of contraception,
National Survey of Family Growth, 2011-2013

	Frequency	Weighted Frequency	Percent	95% Confidence Limits
Age				
15-19	183	1,740,000	43.5	35.98 - 51.10
20-29	919	9,341,000	56.6	52.90 - 60.36
30-44	1356	17,342,000	67.2	64.06 - 70.31
Race/ethnicity				
Hispanic	576	5,229,000	57.3	52.92 - 61.64
NH White	1211	17,373,000	64.7	61.28 - 68.10
NH Black	494	3,657,000	56.6	51.09 - 62.16
Marital status				
Married	941	13,629,000	70.9	67.86 - 73.98
Cohab	402	4,481,000	62.4	56.26 - 68.58
Wid/div/sep	335	3,173,000	62.1	56.25 - 67.90
Never married	780	7,139,000	48.3	43.69 - 52.81
Federal poverty level				
<100	825	7,335,000	57.6	53.78 - 61.42
100-199	555	6,015,000	60.6	56.31 - 64.88

200-399	656	8,608,000	63.8	59.84 - 67.69
400-499	152	2,462,000	66.6	57.65 - 75.45
500+	270	4,001,000	62.1	54.83 - 69.41

* Women are considered to be at risk of unintended pregnancy if they are not pregnant, not seeking pregnancy, are fecund, and have ever had sex.

2b4.3. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors (clinical factors or sociodemographic factors) used in the statistical risk model or for stratification by risk (*e.g., potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of p<0.10; correlation of x or higher; patient factors should be present at the start of care*)

We recommended stratifying the client population by age so that rates for adolescents can be tracked separately from those for adult women. We propose this stratification for purposes of QI but not as a method of risk-adjustment. Teen pregnancy is worthy of a separate focus because of the large potential negative impact on the life of the teen and her child(ren), and the existence of unique programs and contraceptive counseling approaches tailored to this population.

2b4.4a. What were the statistical results of the analyses used to select risk factors?

Not applicable.

2b4.4b. Describe the analyses and interpretation resulting in the decision to select SDS factors (e.g. prevalence of the factor across measured entities, empirical association with the outcome, contribution of unique variation in the outcome, assessment of between-unit effects and within-unit effects)

Not applicable.

2b4.5. Describe the method of testing/analysis used to develop and validate the adequacy of the statistical model or stratification approach (*describe the steps—do not just name a method; what statistical analysis was used*)

Provide the statistical results from testing the approach to controlling for differences in patient characteristics (case mix) below.
If stratified, skip to 2b4.9

2b4.6. Statistical Risk Model Discrimination Statistics (*e.g., c-statistic, R-squared*)**:**

Not applicable.

2b4.7. Statistical Risk Model Calibration Statistics (*e.g., Hosmer-Lemeshow statistic*):

Not applicable.

2b4.8. Statistical Risk Model Calibration – Risk decile plots or calibration curves:

Not applicable.

2b4.9. Results of Risk Stratification Analysis:

Not applicable.

2b4.10. What is your interpretation of the results in terms of demonstrating adequacy of controlling for differences in patient characteristics (case mix)? (i*.e., what do the results mean and what are the norms for the test conducted*)

Not applicable.

2b4.11. Optional Additional Testing for Risk Adjustment (*not required, but would provide additional support of adequacy of risk model, e.g., testing of risk model in another data set; sensitivity analysis for missing data; other methods that were assessed*)

Not applicable.

_____

2b5. IDENTIFICATION OF STATISTICALLY SIGNIFICANT & MEANINGFUL DIFFERENCES IN PERFORMANCE

2b5.1. Describe the method for determining if statistically significant and clinically/practically meaningful differences in performance measure scores among the measured entities can be identified (*describe the steps—do not just name a method; what statistical analysis was used? Do not just repeat the information provided related to performance gap in 1b)*

Due to the fact that our dataset represents a census of all claims available, rates are assumed to reflect 'true' rates by unit for the data year. Non-sampling error (such as coding or measurement error) is not estimable given our limited access to the claims data and processes. Thus we do not present any confidence intervals for inferential testing results. These assumed-true differences in rates must therefore be evaluated based on practical or clinically meaningful impact.

2b5.2. What were the statistical results from testing the ability to identify statistically significant and/or clinically/practically meaningful differences in performance measure scores across measured entities? (e.g., *number and percentage of entities with scores that were statistically significantly different from mean or some benchmark, different from expected; how was meaningful difference defined*)

The table below summarizes rates at each level. As noted above, since our data contain the entirety of the defined population, estimation of sampling error and related inferential statistics such as confidence intervals are not applicable. More detailed information about the variation of rates by unit within each level can be found in Tables 1-3, which are appended at the end of this document.

Level	Age group	Rate (use of most and moderately effective methods)
Affiliate (PPFA), n=25 Mean (range)	15-20	.73 (.37-.90)
	21-44	.66 (.28-.84)
	15-44	.68 (.31-.86)
Health center (PPFA), n=363 Mean (range)	15-20	.73 (.00-1.0)
	21-44	.66 (.00-1.0)
	15-44	.68 (.00-1.0)
Public health region (IME) Mean (range)	15-20	.62 (.52-.67)
	21-44	.60 (.56-.65)
	15-44	.60 (.58-.66)
Benefit type (IME) Mean (range)	15-20	.62 (.40-.79)
	21-44	.60 (.28-.73)
	15-44	.60 (.32-.74)
PH Region by benefit type (IME) Mean (range)	15-20	.62 (.39-.84)
	21-44	.60 (.27-.74)
	15-44	.60 (.31-.78)

Health plan/HMO (WMP) Mean (range)	15-20	.46 (.42-.50)
	21-44	.38 (.34-.43)
	15-44	.40 (.36-.45)

2b5.3. What is your interpretation of the results in terms of demonstrating the ability to identify statistically significant and/or clinically/practically meaningful differences in performance across measured entities? (i.*e., what do the results mean in terms of statistical and meaningful differences?*)

There are very large and meaningful differences in rates across almost all levels. These differences suggest that it will be possible to identify meaningful differences in performance across measured entities. It also reinforces that there is substantial room for improvement in measure scores.

As more experience is gained from using the measures in different programmatic contexts (e.g., in programs focused on reproductive health services versus general health care providers), it will be possible to recommend benchmarks for the different programmatic contexts. These benchmarks may also take into account subject matter expertise, economic costs, risks of maternal health, pregnancy or birth outcomes, and other contextual criteria. If the measure is endorsed by NQF, OPA expects to convene an expert panel within 2 years to identify appropriate criteria and apply those criteria for the development of recommended benchmarks.

2b6. COMPARABILITY OF PERFORMANCE SCORES WHEN MORE THAN ONE SET OF SPECIFICATIONS
***If only one set of specifications, this section can be skipped*.**

Not applicable.

Note: *This item is directed to measures that are risk-adjusted (with or without SDS factors)* ***OR*** *to measures with more than one set of specifications/instructions (e.g., one set of specifications for how to identify and compute the measure from medical record abstraction and a different set of specifications for claims or eMeasures). It does not apply to measures that use more than one source of data in one set of specifications/instructions (e.g., claims data to identify the denominator and medical record abstraction for the numerator).* ***Comparability is not required when comparing performance scores with and without SDS factors in the risk adjustment model. However, if comparability is not demonstrated for measures with more than one set of specifications/instructions, the different specifications (e.g., for medical records vs. claims) should be submitted as separate measures.***

2b6.1. Describe the method of testing conducted to compare performance scores for the same entities across the different data sources/specifications (*describe the steps—do not just name a method; what statistical analysis was used*)

Not applicable.

2b6.2. What were the statistical results from testing comparability of performance scores for the same entities when using different data sources/specifications? (*e.g., correlation, rank order*)

Not applicable.

2b6.3. What is your interpretation of the results in terms of the differences in performance measure scores for the same entities across the different data sources/specifications? (i.*e., what do the results mean and what are the norms for the test conducted*)

Not applicable.

2b7. MISSING DATA ANALYSIS AND MINIMIZING BIAS

2b7.1. Describe the method of testing conducted to identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias (*describe the steps—do not just name a method; what statistical analysis was used*)

The data source for this measure is claims data. Due to the nature of claims data (i.e., for billing purposes), there is typically very little missing data; further, it is difficult to ascertain when claims data is or is not missing.

2b7.2. What is the overall frequency of missing data, the distribution of missing data across providers, and the results from testing related to missing data? (*e.g., results of sensitivity analysis of the effect of various rules for missing data/nonresponse; if no empirical sensitivity analysis, identify the approaches for handling missing data that were considered and pros and cons of each*)

Not addressed due to the nature of claims data.

2b7.3. What is your interpretation of the results in terms of demonstrating that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias**?** (i*.e., what do the results mean in terms of supporting the selected approach for missing data and what are the norms for the test conducted; if no empirical analysis, provide rationale for the selected approach for missing data*)

Not applicable.

References

Adams JL, Mehrotra A, Thomas JW, McGlynn EA (2010). Physician cost profiling-reliability and risk of misclassification. New England Journal of Medicine; 362:1014–1021.

Adams JL (2009). The Reliability of Provider Profiling: A Tutorial. Rand Health, Santa Monica, CA. Last accessed online at: http://www.rand.org/pubs/technical_reports/TR653.

Eijkenaar F, van Vliet RC. (2013). Profiling individual physicians using administrative data from a single insurer: variance components, reliability, and implications for performance improvement efforts. Med Care. 51(8):731-9

Ene, M, Leighton, EA, Blue, GL, Bell, BA (2015). Models for categorical data using SAS PROC GLIMMIX. Sas Global Forum, Dallas TX, April 26-29, Paper 3430. Last accessed online at: http://support.sas.com/resources/papers/proceedings15/3430-2015.pdf.

Fung, V, Schmittdiel, JA, Fireman, B, Meer, A, Thomas, S, Hsu, J, Selby, JV (2010). Meaningful variation in performance: A systematic literature review. Medical Care, 48(2): 140-148.

He, Y, Selck, F, and Normand, SL (2014). On the accuracy of classifying hospitals on their performance measures. Statistics in Medicine; 33: 1081-1103.

Martin JA, Hamilton BE, Osterman MJ, Curtin SC, Matthews TJ. (2015) Births: final data for 2013. Natl Vital Stat Rep; 64(1):1-65.

Molenberghs G, Laenen A, Vengeneugden T (2007). Estimating reliability and generalizability from hierarchical biomedical data. Journal of Biopharmaceutical Statistics, 17(4): 595-627.

National Committee for Quality Assurance (NCQA). HEDIS 2015: Healthcare Effectiveness Data and Information Set. Vol. 1, narrative. Washington (DC): National Committee for Quality Assurance (NCQA); Retrieved from: http://www.qualitymeasures.ahrq.gov/content.aspx?id=48621

NQF 2009. National Quality Forum, Health Information Technology Expert Panel II- Health IT Enablement of Quality Measurement. Washington DC.

Raudenbush SW and Bryk AS. *Hierarchical Linear Models: Applications and Data Analysis Methods (2nd Edition)*. Sage Publications, 2002: Thousand Oaks.

Ridout MS, Demetrio CGB, Firth D (1999). Estimating intraclass correlation for binary data. Biometrics, 55(1): 137-148.

Romero L, Pazol K, Warner L, Gavin L, Moskosky S, Besera G, et al. (2015). Vital Signs: Trends in Use of Long-Acting Reversible Contraception Among Teens Aged 15–19 Years Seeking Contraceptive Services — United States, 2005–2013. MMWR Weekly Report April 10, 2015 / 64(13);363-369

Scholle SH, Roski J, Adams JL, Dunn DL, Kerr EA, Pilleterre-Dugan D, Jensen RE (2008). Benchmarking physician performance: reliability of individual and composite measures. Am J of Managed Care, 14(12): 833-838.

Sedgh G, Finer LB, Bankole A, Eilers MA, Singh S. (2015) Adolescent pregnancy, birth, and abortion rates across countries: levels and recent trends. J Adolesc Health; 56(2):223-30.

Woltman H, Feldstain A, MacKay JC, Rocchi M. (2012). An introduction to hierarchical linear modeling. Tutorials in Quantitative Methods for Psychology; 8(1):52–69.

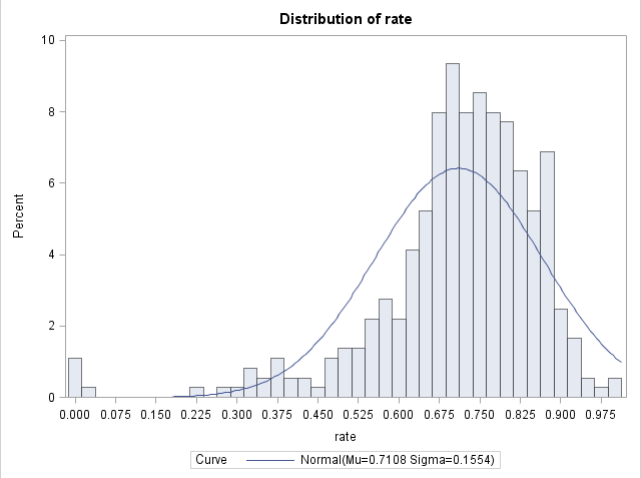
Wu S, Crespi CM, Wong WK (2012). Comparison of methods for estimating the intraclass correlation coefficient for binary responses in cancer prevention cluster randomized trials. Contemp Clin Trials, 33(5): 869-880.

Table 1. Rates and reliabilities for moderate or most effective contraceptive method, 25 PPFA affiliates, 2014.

Affiliate ID	MEM_MethodMost: 15 to <21 Years					MEM_MethodMost: 21 to 45 years					MEM_MethodMost: all age groups				
		Used Most/Mod	Total N	Rate	HC Within Affiliate Reliability		Used Most/Mod	Total N	Rate	HC Within Affiliate Reliability		Used Most/Mod	Total N	Rate	HC Within Affiliate Reliability
1		6765	7869	0.860	**0.9822**		23438	29638	0.791	**0.9990**		30203	37507	0.805	**0.9993**
3		19611	26591	0.738	**0.9850**		63945	88881	0.719	**0.9984**		83556	115472	0.724	**0.9981**
4		2713	4147	0.654	**0.6096**		12900	21430	0.602	**0.9597**		15613	25577	0.610	**0.9630**
5		33452	42698	0.783	**0.9910**		92727	131187	0.707	**0.9955**		126179	173885	0.726	**0.9971**
6		1710	2651	0.645	**0.8348**		4390	7362	0.596	**0.9469**		6100	10013	0.609	**0.9563**
9		17261	25268	0.683	**0.9852**		58446	88455	0.661	**0.9936**		75707	113723	0.666	**0.9952**
10		11239	15188	0.740	**0.9090**		30853	47698	0.647	**0.9281**		42092	62886	0.669	**0.9637**
12		4355	4839	0.900	**0.8011**		8599	10209	0.842	**0.9419**		12954	15048	0.861	**0.9604**
37		1318	1965	0.671	**0.6152**		2187	4194	0.521	**0.6709**		3505	6159	0.569	**0.7056**
38		4071	6093	0.668	**0.9502**		5683	10645	0.534	**0.9529**		9754	16738	0.583	**0.9743**
40		3628	5030	0.721	**0.9150**		5991	10843	0.553	**0.9264**		9619	15873	0.606	**0.9544**
41		4130	5466	0.756	**0.9512**		11021	17562	0.628	**0.8917**		15151	23028	0.658	**0.9620**
44		9903	11489	0.862	**0.9647**		27664	33620	0.823	**0.9848**		37567	45109	0.833	**0.9879**
47		4383	5644	0.777	**0.9793**		10699	16648	0.643	**0.9899**		15082	22292	0.677	**0.9946**
53		5136	8741	0.588	**0.9696**		12678	28791	0.440	**0.9691**		17814	37532	0.475	**0.9850**
54		2277	3122	0.729	**0.8237**		3636	6614	0.550	**0.9522**		5913	9736	0.607	**0.9765**
59		2851	3682	0.774	**0.9903**		6470	9778	0.662	**0.9991**		9321	13460	0.692	**0.9994**
60		258	436	0.592	**0.0000**		599	1265	0.474	**0.0000**		857	1701	0.504	**0.0000**
70		1944	4154	0.468	**0.9985**		3507	12436	0.282	**0.9996**		5451	16590	0.329	**0.9997**
73		406	996	0.408	**0.9863**		789	2825	0.279	**0.9874**		1195	3821	0.313	**0.9933**
75		431	1171	0.368	**0.9828**		1605	5070	0.317	**0.9946**		2036	6241	0.326	**0.9958**
76		1737	3817	0.455	**0.9368**		4819	11648	0.414	**0.8733**		6556	15465	0.424	**0.9425**
77		8635	11359	0.760	**0.9693**		21324	31393	0.679	**0.9655**		29959	42752	0.701	**0.9833**
79		721	1260	0.572	**0.8878**		2162	5149	0.420	**0.9440**		2883	6409	0.450	**0.9625**
81		182	294	0.619	**0.0000**		806	1561	0.516	**0.0000**		988	1855	0.533	**0.0000**
Total or Mean		**149117**	**203970**	**0.731**			**416938**	**634902**	**0.657**			**566055**	**838872**	**0.675**	
			σ Level 2	**ICC**	**Overall Affiliate Reliability**			**σ Level 2**	**ICC**	**Overall Affiliate Reliability**			**σ Level 2**	**ICC**	**Overall Affiliate Reliability**
Reliability using Median Affiliate Patient Volume	**Median n**	**4839**	**0.4334**	**0.1164**	**0.9984**	**Median n**	**11648**	**0.4624**	**0.1232**	**0.9994**	**Median n**	**16590**	**0.4448**	**0.1191**	**0.9996**
Reliability using Minimum Patient Volume (Floor)	Min n	294	0.4334	0.1164	0.9748	Min n	1265	0.4624	0.1232	0.9944	Min n	1701	0.4448	0.1191	0.9957

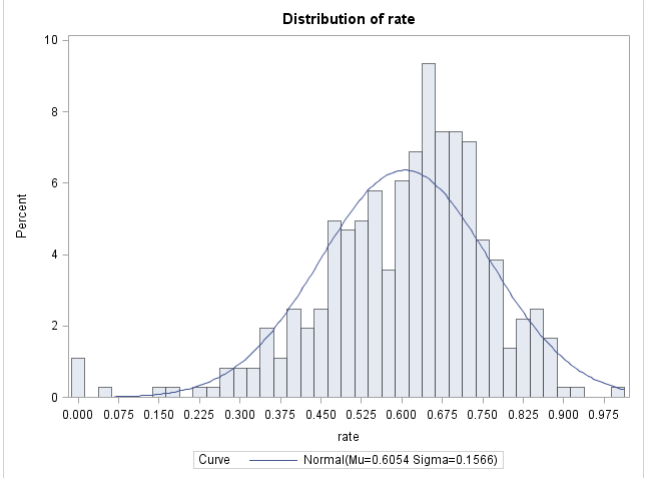
Table 2. Distributions of rates and ICCs among health centers (n=363) for moderate/most effective methods among 25 PPFA affiliates, 2014

MEM_MethodMost: 15 to <21 Years	MEM_MethodMost: 21 to 45 years	ME



	cases (n)	rate
Mean	561.9	0.71
Median	366	0.73
SD	552.3	0.16
Variance	305043	0.02
Range	2976	1.00
Interquartile	629	0.15
Median ICC		0.06

Quantile	cases (n)	rate
100% Max	2984	1.00
95%	1766	0.90
90%	1410	0.87
75% Q3	787	0.81
50% Med	366	0.73
25% Q1	158	0.66
10%	83	0.54
5%	53	0.41
0% Min	8	0.00



	cases (n)	rate
Mean	1749.0	0.61
Median	1016	0.63
SD	1909	0.16
Variance	3645550	0.02
Range	11360	1.00
Interquartile	2145	0.19
Median ICC		0.05

Quantile	cases (n)	rate
100% Max	11391	1.00
95%	5489	0.83
90%	4544	0.78
75% Q3	2516	0.71
50% Med	1016	0.63
25% Q1	371	0.52
10%	149	0.41
5%	92	0.34
0% Min	31	0.00

Table 3. Rates and reliabilities for moderate or most effective contraceptive method, Iowa Medicaid Enterprise, 2013, by region and type of benefit

Public Health Region	MEM_MethodMost: 15 to <21 Years						MEM_MethodMost: 21 to 45 years						MEM_MethodMost: all age groups					
	Not Used	Used Most/ Mod	Total N	Rate	VarL1	ICC	Region Reliability (VarL1)	Not Used	Used Most/ Mod	Total N	Rate	VarL1	ICC	Region Reliability (VarL1)	Not Used	Used Most/ Mod	Total N	Rate	VarL1	ICC	Region Reliability (VarL1)
1	1384	2076	3460	0.600				37501	5838	9588	0.609				51341	7914	13048	0.607			
2	3800	774	1154	0.671				10131	1893	2906	0.651				13931	2667	4060	0.657			
3	565	611	1176	0.520				26112	1914	3175	0.603				18261	2525	4351	0.580			
4	415	672	1087	0.618				26218	1625	2887	0.563				16772	2297	3974	0.578			
5	572	1129	1701	0.664				90407	2469	4359	0.566				24625	3598	6060	0.594			
6	1148	1973	3121	0.632				9	6057	10136	0.598				22717	8030	13257	0.606			
Total or Mean	**4464**	**7235**	**11699**	**0.618**				**13255**	**19796**	**33051**	**0.599**				**17719**	**27031**	**44750**	**0.604**			
Median Patient Volume Among Affiliates		Median n	1438.5	0.04018	0.0121	**0.9461**			Median n	3767	0.01365	0.0041	**0.9399**			Median n	5205.5	0.01109	0.00034	**0.9461**	
Minimum Patient Volume (Floor)		Min n	1087	0.04018	0.0121	0.9299			Min n	2887	0.01365	0.0041	0.9229			Min n	3974	0.01109	0.00034	0.9305	

Type of benefit (family planning waiver vs general Medicaid)	MEM_MethodMost: 15 to <21 Years				MEM_MethodMost: 21 to 45 years				MEM_MethodMost: all age groups			
	Not Used	Used Most/ Mod	Total N	Rate	Not Used	Used Most/ Mod	Total N	Rate	Not Used	Used Most/ Mod	Total N	Rate
Family Planning Waiver	1333	5112	6445	0.793	6430	17138	23568	0.727	7763	22250	30013	0.741
General Medicaid	3131	2123	5254	0.404	6825	2658	9483	0.280	9956	4781	14737	0.324
Total or Mean	**4464**	**7235**	**11699**	**0.618**	**13255**	**19796**	**33051**	**0.599**	**17719**	**27031**	**44750**	**0.6**

				VarL2	ICC	Type of Benefit Reliability (VarL2)				VarL2	ICC	Type of Benefit Reliability (VarL2)				VarL2	ICC	Type of Benefit Reliability (VarL2)
Reliability Based on Median Patient Volume Among Health Centers		Median n	5849.5	0.4778	0.12688	**0.9988**		Median n	16525.5	0.01878	0.00577	**0.98995**		Median n	22375	0.15988	0.04063	**0.9991**
Calculated Based on Minimum Patient Volume (Floor)		Min n	5254	0.4778	0.12688	0.9987		Min n	9483	0.01878	0.00577	0.98919		Min n	4444	0.15988	0.04063	0.9986
Region 1/Family Planning Waiver	941	595	1536	0.387			1801	743	2544	0.292			2742	1338	4080	0.3288		
Region 1/ General Medicaid	443	1481	1924	0.770			1949	5095	7044	0.723			2392	6576	8968	0.7333		
Region 2/Family Planning Waiver	271	192	463	0.415			477	178	655	0.272			748	370	1118	0.3310		
Region 2/ General Medicaid	109	582	691	0.842			536	1715	2251	0.762			645	2297	2942	0.7810		
Region 3/Family Planning Waiver	405	204	609	0.335			535	240	775	0.310			940	444	1384	0.3211		
Region 3/ General Medicaid	160	407	567	0.718			726	1674	2400	0.698			886	2081	2967	0.7010		
Region 4/Family Planning Waiver	297	239	536	0.446			640	261	901	0.290			937	500	1437	0.3488		
Region 4/ General Medicaid	118	433	551	0.786			622	1364	1986	0.687			740	1797	2537	0.7080		
Region 5/Family Planning Waiver	404	337	741	0.455			1089	408	1497	0.273			1493	745	2238	0.3330		
Region 5/ General Medicaid	168	792	960	0.825			801	2061	2862	0.720			969	2853	3822	0.7460		
Region 6/Family Planning Waiver	813	556	1369	0.406			2283	828	3111	0.266			3096	1384	4480	0.3090		
Region 6/	335	1417	175	0.809			1796	5229	7025	0.74			2131	6646	8777	0.		

General Medicaid			2						4						7			
Total or Mean	**4464**	**7235**	**11699**	**0.618**			**13255**	**19796**	**33051**	**0.599**			**17719**	**27031**	**44750**	**0.604**		
				VarL2	**ICC**	Type of Benefit by Public Health Region Reliability (Var L2)				**VarL2**	**ICC**	Type of Benefit by Public Health Region Reliability (VarL2)				**VarL2**	**ICC**	Type of Benefit by Public Health Region Reliability (Var L2)
Reliability Based on Median Patient Volume Among Health Centers		Median n	716	0.7862	0.1929	**0.9942**		Median n	2325.5	0.89898	0.2148	**0.99984**		Median n	2954.5	0.782	0.19220	**0.9986**
Calculated Based on Minimum Patient Volume (Floor)		Min n	463	0.7862	0.1929	0.9910		Min n	655	0.89898	0.2148	0.99444		Min n	1118	0.782	0.19220	0.9963

Table 4. Rates and reliabilities for moderate or most effective contraceptive method, Wisconsin Medicaid, 2014, by health plan

HMO	MEM_MethodMost: 15 to <21 Years						MEM_MethodMost: 21 to 45 years						MEM_MethodMost: all age groups					
	Not Used	Used Most/ Mod	Total N	Rate	VarL1	ICC	Overall HMO Reliability (Var L1)	Not Used	Used Most/ Mod	Total N	Rate	VarL1	ICC	Overall HMO Reliability (Var L1)	Not Used	Used Most/ Mod	Total N	Rate	VarL1	ICC	Overall HMO Reliability (Var L1)
1	2827	2005	4832	0.415				8462	5581	140043	0.397				112897	7586	18875	0.402			
2	972	866	1838	0.471				3560	2128	5688	0.374				45322	2994	7526	0.398			
3	490	430	920	0.467				1852	1010	2862	0.353				23421	1440	3782	0.381			
4	924	871	1795	0.485				3471	2210	5681	0.389				43953	3081	7476	0.412			
5	649	582	1231	0.473				2614	1322	3936	0.336				32631	1904	5167	0.368			
6	113	106	219	0.484				451	274	725	0.378				564	380	944	0.403			
7	281	277	558	0.496				1032	576	1608	0.358				1313	853	2166	0.394			
8	178	174	352	0.494				663	433	1096	0.395				841	607	1448	0.419			
9	930	693	1623	0.427				4019	2145	6164	0.348				49492	2838	7787	0.364			
10	331	287	618	0.464				1023	660	1683	0.392				1354	947	2301	0.412			
11	2656	2242	4898	0.458				9082	6084	151616	0.401				117378	8326	20064	0.415			
12	660	579	1239	0.467				2776	1514	4290	0.353				3436	2093	5529	0.379			
13	146	123	269	0.457				517	336	853	0.394				663	459	1122	0.409			
14	1158	991	2149	0.461				3692	1904	5596	0.340				4850	2895	7745	0.374			
15	29	27	56	0.482				158	82	240	0.342				187	109	296	0.368			
16	2748	2366	5114	0.463				11763	7112	18875	0.377				145119	9478	23989	0.395			
17	283	276	559	0.494				879	654	1533	0.427				1162	930	2092	0.445			
Total or Mean	**15375**	**12895**	**28270**	**0.456**				**56014**	**34025**	**90039**	**0.378**				**71389**	**46920**	**118309**	**0.397**			
Median Patient Volume Among Affiliates		Median n	1231	0.005593	0.0017	**0.6767**			Median n	3936	0.009698	0.029	**0.9206**			Median n	5167	0.006053	0.018	**0.9048**
Minimum Patient Volume (Floor)		Min n	56	0.005593	0.0017	0.0869			Min n	240	0.009698	0.029	0.4143			Min n	296	0.006053	0.018	0.3526

3. Feasibility

Extent to which the specifications including measure logic, require data that are readily available or could be captured without undue

burden and can be implemented for performance measurement.

3a. Byproduct of Care Processes

For clinical measures, the required data elements are routinely generated and used during care delivery (e.g., blood pressure, lab test, diagnosis, medication order).

3a.1. **Data Elements Generated as Byproduct of Care Processes.**

Coded by someone other than person obtaining original information (e.g., DRG, ICD-9 codes on claims)

If other:

3b. Electronic Sources

The required data elements are available in electronic health records or other electronic sources. If the required data are not in electronic health records or existing electronic sources, a credible, near-term path to electronic collection is specified.

3b.1. **To what extent are the specified data elements available electronically in defined fields?** *(i.e., data elements that are needed to compute the performance measure score are in defined, computer-readable fields)*

ALL data elements are in defined fields in electronic claims

3b.2. **If ALL the data elements needed to compute the performance measure score are not from electronic sources, specify a credible, near-term path to electronic capture, OR provide a rationale for using other than electronic sources.**

3b.3. **If this is an eMeasure, provide a summary of the feasibility assessment in an attached file or make available at a measure-specific URL.**

Attachment:

3c. Data Collection Strategy

Demonstration that the data collection strategy (e.g., source, timing, frequency, sampling, patient confidentiality, costs associated with fees/licensing of proprietary measures) can be implemented (e.g., already in operational use, or testing demonstrates that it is ready to put into operational use). For eMeasures, a feasibility assessment addresses the data elements and measure logic and demonstrates the eMeasure can be implemented or feasibility concerns can be adequately addressed.

3c.1. **Describe what you have learned/modified as a result of testing and/or operational use of the measure regarding data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, time and cost of data collection, other feasibility/implementation issues.**

IF a PRO-PM, consider implications for both individuals providing PROM data (patients, service recipients, respondents) and those whose performance is being measured.

The measure has been extensively piloted in several programmatic contexts (Planned Parenthood, two state Medicaid programs), and is being reported by 13 state Medicaid programs that are funded by the Center for Medicaid and CHIP Services (CMCS) to report from 2015-2018 on the measure as part of the Maternal and Infant Health initiative. Overall, these experiences have confirmed that the measures can be feasibly calculated using existing claims data. Some other lessons learned are summarized below:

- The use of claims data results in a sample that is nearly identical to use of the birth record. There were 14,757 live births identified in the Medicaid claims data using the SAS code written for the measure specifications and 15,212 Medicaid live births were linked to the birth certificate in 2013, which is only a difference of 455 live births (i.e., 3%).
- State-specific NDC codes may need to be added: Several of the CMCS-funded states found that some NDC codes were missing from the measure specifications, especially for oral contraception. Obtaining a comprehensive list of NDC codes for contraception is challenging, and we relied on the list compiled by NCQA for the calculation of the HEDIS chlamydia screening measure. We have asked states to include the additional NDC codes when they calculate the measure, but to report to us the additional codes at the time of submission. This is consistent with the approach used by NCQA for the HEDIS chlamydia screening measure.
- Feasibility and biases introduced by using a 'look back' period. Several states have inquired about the use of a 'look back' period to identify women who were sterilized or received LARC in a year preceding the measurement year, or who had a procedure that caused them to become infecund for non-contraceptive reasons (e.g., oophorectomy). We are open to exploring this, but several challenges related to feasibility and potential biases caused us to decide against including a look-back period in the current version of the measure: (1) sterilization is permanent, so a plan would have to look back through the entire period of time the woman was a client; (2) LARC last for a period 3-10 years, so an extensive period of look-back would be needed for this purpose as

well; and (3) high rates of moving on and off a health plan may make it very hard to interpret the results of a look-back, especially if only a small percentage of currently enrolled women were enrolled for the entire look-back period.

Finally, existing administrative claims data has several limitation with regard to the measurement of unintended pregnancy. In particular, the data does not capture the client's history of sexual experience, their desire to become pregnant, and sterilization or LARC insertion in a year preceding the measurement year. In the medium-term, these limitations are not a serious concern because there is so much room for improvement and the focus will be on demonstrating percentage point increases (e.g., by 15 percentage points over three years). Yet for the longer-term, OPA is actively pursuing efforts to develop either a hybrid or all-electronic version of the measure. A hybrid measure will be piloted in two community health centers in the coming year, and we are working with Integrating the Healthcare Enterprise to create a family planning profile that includes all data elements needed to compute the measures with electronic data. If these plans are successful, we expect to be able to submit an application for either a hybrid or e-measure version of the measure within a 3-year period of time.

3c.2. **Describe any fees, licensing, or other requirements to use any aspect of the measure as specified** *(e.g., value/code set, risk model, programming code, algorithm)*.
Not applicable. The measure specifications, code lists, programming code and NSFG tables needed to interpret scores will all be available at no charge on the OPA website.

4. Usability and Use

Extent to which potential audiences (e.g., consumers, purchasers, providers, policy makers) are using or could use performance results for both accountability and performance improvement to achieve the goal of high-quality, efficient healthcare for individuals or populations.

4a. Accountability and Transparency

Performance results are used in at least one accountability application within three years after initial endorsement and are publicly reported within six years after initial endorsement (or the data on performance results are available). If not in use at the time of initial endorsement, then a credible plan for implementation within the specified timeframes is provided.

4.1. Current and Planned Use

NQF-endorsed measures are expected to be used in at least one accountability application within 3 years and publicly reported within 6 years of initial endorsement in addition to performance improvement.

Planned	Current Use (for current use provide URL)
Public Reporting	Quality Improvement (Internal to the specific organization) https://www.medicaid.gov/Medicaid-CHIP-Program-Information/By-Topics/Quality-of-Care/Maternal-and-Infant-Health-Care-Quality.html CMCS Maternal and Infant Health Initiative OPA Title X family planning program www.familyplanningdashboard.com

4a.1. For each CURRENT use, checked above, provide:

- Name of program and sponsor
- Purpose
- Geographic area and number and percentage of accountable entities and patients included

The Center for Medicaid and CHIP Services

As part of the CMCS Maternal and Infant Health Initiative, state Mediciad programs will report on the measure on a voluntary basis. Nearly two out of every three adult women enrolled in Medicaid are in their reproductive years (ages 19-44), and Medicaid currently finances about 45% of all births in the United States. Recognizing this, the Centers for Medicaid and CHIP Services (CMCS) has developed the Maternal and Infant Health initiative to improve the quality of maternity care, birth outcomes and in measuring how care is delivered to women. CMCS will collect and report data on the new contraceptive use measure on a developmental basis, to help states track the use of most and moderately effective methods of contraception and to drive changes in care practices and delivery. The measure specifications developed for use in the Maternal and Infant Health initiative are found in Appendix C. For more information about the Maternal and Infant Health initiative, see: https://www.medicaid.gov/Medicaid-CHIP-Program-Information/By-Topics/Quality-of-Care/Maternal-and-Infant-Health-Care-Quality.html.

OPA's Title X Program

In 2014, the Title X program funded 94 grantees that support a network of 4100 service sites, which in turn served 4.1 million clients. Approximately 75% of US counties have at least one Title X-funded clinic. In addition, in 2015-2016, OPA has developed an online contraceptive measures calculator for use throughout its Title X program. The calculator enables grantees, using FPAR data, to automatically calculate the two measures for each site in its network. OPA is also piloting a learning collaborative, based on the Institute of Healthcare Improvement's Breakthrough Series model, to support grantees to assess and optimize their performance on the access and outcome measures through employing strategies documented in an evidence-based change package. The learning collaborative involves coaching and supporting the members through the plan, do, study, act cycle for selected change package strategies. The collaborative includes an on-line community of practice to further promote peer exchange and learning. The collaborative serves as a model that can be spread throughout the Title X program and adapted in other settings. The change package and other materials from the collaborative, once tested, will be made available for use by the full reproductive health community. As an example, the Iowa Department of Public Health (IDPH) is a Title X grantee that has pioneered use of this performance measure within its service site network, and provided the following comments on the usability of the measure: "IDPH has used the data to assess access to most and moderately effective contraceptive methods in the Title X provider network; explored barriers/challenges and opportunities to improve access; identified training needs for grantees and strategies to improve performance. The data has provided IDPH a standardized way of looking at access to contraceptive methods, how services are being provided and how to identify gaps in services. The measure will give us an opportunity to consider why we are seeing the results we

are seeing, to think about those results and determine if and where changes can be made to improve outcomes. We have used the data to help develop the competitive application, and will use it in years 2-5 of the Title X award period as a variable in the funding formula."

Planned Parenthood Federation of America
In 2014, Planned Parenthood of America (PPFA) comprised 66 independently incorporated affiliates, operating approximately 700 health centers in the United States, providing reproductive health care to nearly 2.7 million patients. In August 2012, Planned Parenthood Federation of America (PPFA) added a Clinical Quality Improvement (CQI) Department to its Health Care Division. The CQI Department coordinates a federation-wide clinical quality improvement program working with Affiliates. As part of the federation-wide quality improvement program and to capitalize on this investment in health IT, a set of core reports were built as key measures of quality of care and health outcomes including use of contraception services. Nearly 70% of the affiliates partner with the PPFA CQI Department and receive quality reports on key clinical measures as well as technical assistance for quality improvement activities.

4a.2. **If not currently publicly reported OR used in at least one other accountability application (e.g., payment program, certification, licensing) what are the reasons?** (*e.g., Do policies or actions of the developer/steward or accountable entities restrict access to performance results or impede implementation?*)

4a.3. **If not currently publicly reported OR used in at least one other accountability application, provide a credible plan for implementation within the expected timeframes -- any accountability application within 3 years and publicly reported within 6 years of initial endorsement.** (*Credible plan includes the specific program, purpose, intended audience, and timeline for implementing the measure within the specified timeframes. A plan for accountability applications addresses mechanisms for data aggregation and reporting.*)
1) The Measure Applications Partnership (MAP) recommended inclusion of these measures in the core adult and child measure set, pending NQF endorsement.
2) OPA's Title X program intends to make publicly available the performance of all grantees (individually and aggregated by state) within a 3 year period of time.

4b. Improvement

Progress toward achieving the goal of high-quality, efficient healthcare for individuals or populations is demonstrated. If not in use for performance improvement at the time of initial endorsement, then a credible rationale describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.

4b.1. **Progress on Improvement. (Not required for initial endorsement unless available.)**
Performance results on this measure (current and over time) should be provided in 1b.2 and 1b.4. Discuss:

- **Progress (trends in performance results, number and percentage of people receiving high-quality healthcare)**
- **Geographic area and number and percentage of accountable entities and patients included**

Not applicable

4b.2. **If no improvement was demonstrated, what are the reasons? If not in use for performance improvement at the time of initial endorsement, provide a credible rationale that describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.**
Not applicable

4c. Unintended Consequences

The benefits of the performance measure in facilitating progress toward achieving high-quality, efficient healthcare for individuals or populations outweigh evidence of unintended negative consequences to individuals or populations (if such evidence exists).

4c.1. **Were any unintended negative consequences to individuals or populations identified during testing; OR has evidence of unintended negative consequences to individuals or populations been reported since implementation? If so, identify the negative unintended consequences and describe how benefits outweigh them or actions taken to mitigate them.**
No unintended negative consequences were identified. The one issue that has arisen as a potential concern is that the measure may lead to coercive practices in which women are not offered a free choice of methods and are pressured into using the most and moderately effective methods (Gold 2014, Dehlendorf 2015). We do not think this will be a concern for several reasons:

- Nine methods of contraception are included in the numerator, which are treated as being of equal value when calculating the measure. Hence, there is a wide range of client choice and the measure is not likely to prevent providers from delivering care in a fully client-centered, non-coercive, manner.

- Although existing research (Harper 2015, Winner 2012) show a high percentage of women will choose LARC when given the opportunity, we have deliberately not set a benchmark until we have more experience using the measure in real-life settings. The lack of a benchmark should lessen pressure on providers to encourage all women to use a most or moderately effective method.

- CDC-OPA recommendations describe in detail how to provide client-centered, non-coercive contraceptive counseling, and efforts to support use of the measure should be accompanied by efforts to increase awareness of the CDC-OPA recommendations (CDC/OPA 2014). Further, OPA has funded the development of training on how to provide client-centered training, which is available to all providers on the OPA-supported training website (www.fpntc.org).

References:

- Dehlendorf C, Bellanca H, Policar M (2015). Performance measures for contraceptive care: what are we actually trying to measure? Contraception. Jun;91(6):433-7.
- Gold R (2014). Guarding Against Coercion While Ensuring Access: A Delicate Balance. Guttmacher Policy Review. Summer 2014, Volume 17, Number 3.
- Harper C, Rocca CH, Thompson KM, Morfesis J, Goodman S, Darney PD, Westhoff CL, Speidel JJ (2015). Reductions in pregnancy rates in the USA with long-acting reversible contraception: a cluster randomised trial. Lancet. Volume 386, No. 9993, p562–568, 8 August 2015.
- Winner B, Peipert J, Qiuhong Z, Buckel C, Madden T et al (2012). Effectiveness of Long-Acting Reversible Contraception, The New England Journal of Medicine, 366 (21): 1998-2007.
- CDC/OPA (2014). Providing Quality Family Planning Services: Recommendations of CDC and the US Office of Population Affairs, MMWR Recommendations and Reports, April 24, 2014.
- OPA-funded training materials are available at this website: www.fpntc.org

5. Comparison to Related or Competing Measures

If a measure meets the above criteria and there are endorsed or new related measures (either the same measure focus or the same target population) or competing measures (both the same measure focus and the same target population), the measures are compared to address harmonization and/or selection of the best measure.

5. Relation to Other NQF-endorsed Measures
Are there related measures (conceptually, either same measure focus or target population) or competing measures (conceptually both the same measure focus and same target population)? If yes, list the NQF # and title of all related and/or competing measures.
No

5.1a. **List of related or competing measures (selected from NQF-endorsed measures)**

5.1b. **If related or competing measures are not NQF endorsed please indicate measure title and steward.**

5a. Harmonization

The measure specifications are harmonized with related measures;
OR
The differences in specifications are justified

5a.1. **If this measure conceptually addresses EITHER the same measure focus OR the same target population as NQF-endorsed measure(s):**
Are the measure specifications completely harmonized?

5a.2. **If the measure specifications are not completely harmonized, identify the differences, rationale, and impact on interpretability and data collection burden.**

5b. Competing Measures

The measure is superior to competing measures (e.g., is a more valid or efficient way to measure);
OR
Multiple measures are justified.

5b.1. **If this measure conceptually addresses both the same measure focus and the same target population as NQF-endorsed measure(s):**
Describe why this measure is superior to competing measures (e.g., a more valid or efficient way to measure quality); OR provide a rationale for the additive value of endorsing an additional measure. (Provide analyses when possible.)
NOTE: OPA is submitting two other applications for NQF endorsement, which are complementary to this measure application. One of the applications focuses on use of most and moderately effective contraceptive methods in a key sub-population of women at risk of unintended pregnancy, i.e., postpartum women. The other application focuses on use of a sub-set of contraceptive methods, i.e., use of long-acting reversible contraception (LARC); the goal of this measure to monitor whether women have access to LARC methods as determined by whether any units report very low levels of LARC use (e.g., less than 1-2 percent) or at a level that is substantially below the mean when compared to other reporting units.

Appendix

A.1 Supplemental materials may be provided in an appendix. All supplemental materials (such as data collection instrument or methodology reports) should be organized in one file with a table of contents or bookmarks. If material pertains to a specific submission form number, that should be indicated. Requested information should be provided in the submission form and required attachments. There is no guarantee that supplemental materials will be reviewed.

Attachment **Attachment:** nqf_Appendix_MOST_MOD.pdf

Contact Information

Co.1 **Measure Steward (Intellectual Property Owner):** US Office of Population Affairs
Co.2 **Point of Contact:** Loretta, Gavin, loretta.gavin@hhs.gov, 240-453-2826-
Co.3 **Measure Developer if different from Measure Steward:** US Office of Population Affairs
Co.4 **Point of Contact:** Loretta, Gavin, loretta.gavin@hhs.gov, 240-453-2826-

Additional Information

Ad.1 **Workgroup/Expert Panel involved in measure development**
Provide a list of sponsoring organizations and workgroup/panel members' names and organizations. Describe the members' role in measure development.
The formal steward of the measure is OPA, but the measure is also supported by the U.S. Centers for Disease Control and Prevention, Division of Reproductive Health (http://www.cdc.gov/reproductivehealth/index.html).
Representatives of numerous organizations were involved in helping to develop the measure. Their roles included helping define the conceptual basis and rationale for the measure, piloting and testing data, reviewing draft measure specifications, and/or reviewing the NQF application.

- U.S. Office of Population Affairs: Susan Moskosky MS
- U.S. Centers for Disease Control and Prevention: Peter Briss MD, Gladys Martinez PhD, Lisa Romero PhD, Cheryl Robbins PhD, Karen Pazol PhD, Maria Rivera MPH
- Planned Parenthood Federation of America: Jennifer Fuld PhD, Carolyn Westhoff MD, Mytri Singh MPH, Joycelyn Benson MPH, Kellan Smith
- U.S. Center for Medicaid and CHIP Services: Lekisha Daniel-Robinson
- Iowa Department of Public Health and Iowa Medicaid Enterprise: Denise Wheeler MS, Debra Kane PhD, Brittni Frederiksen PHD, Mikki Stier, Julie Lovelady, Dr. Jason Kessler, Sally Nadolsky (retired), Mark McMahon, Gerd Clabaugh, Brenda Dobson, Marcus Johnson-Miller, Abigail Holicky, Jessica Riggs
- Wisconsin Department of Health Services: Crystal Gibson MPH, Angela Rohan PhD
- Illinois Medicaid: Julie Doetsch, Gwen Smith
- Far Harbor LLC: Philip A. Hastings, PhD and Prasant Mohanty, MBBS, MPH
- National Contraceptive Quality Measures Workgroup

Measure Developer/Steward Updates and Ongoing Maintenance
Ad.2 **Year the measure was first released:**
Ad.3 **Month and Year of most recent revision:**
Ad.4 **What is your frequency for review/update of this measure?**
Ad.5 **When is the next scheduled review/update for this measure?**

Ad.6 **Copyright statement:** Not applicable
Ad.7 **Disclaimers:** Not applicable

Ad.8 **Additional Information/Comments:** Not applicable

MEASURE WORKSHEET

This document summarizes the evaluation of the measure as it progresses through NQF's Consensus Development Process (CDP). The information submitted by measure developers/stewards is included after the Brief Measure Information, Preliminary Analysis, and Pre-meeting Public and Member Comments sections.

To navigate the links in the worksheet: Ctrl + click link to go to the link; ALT + LEFT ARROW to return

Brief Measure Information

NQF #: 2904
Measure Title: Contraceptive Care - Access to LARC
Measure Steward: US Office of Population Affairs
Brief Description of Measure: Percentage of women aged 15-44 years at risk of unintended pregnancy that is provided a long-acting reversible method of contraception (i.e., implants, intrauterine devices or systems (IUD/IUS).
It is an access measure because it is intended to identify situations in which women do not have access to the long-acting reversible methods of contraception (LARC), i.e., contraceptive implants and intrauterine devices.
Developer Rationale: Contraception is a highly effective clinical preventive service intended to prevent teen and unintended pregnancy and space births. The type of contraceptive method used by a woman is strongly associated with her risk of unintended pregnancy [1, 2]. The most effective methods (sterilization and the long-acting reversible methods of IUDs and implants) have a failure rate that is less than 1% per year under typical use; the moderately effective methods (shot, pill, patch, ring and diaphragm) have a typical failure rate of 6-12% per year; the least effective methods have a typical failure rate of 18-28%; and 85% of women will become pregnant in a year if they do not use any contraceptive method at all [2]. The proposed performance measure is based on the fact that the long-acting reversible methods of contraception (LARC), i.e., IUDs and implants, are substantially more effective than other reversible methods at preventing unintended pregnancy, and is designed to ensure women have access to these methods.

Existing research shows that many women will choose LARC methods if given the opportunity to do so. For example, in a recent large prospective study, almost three-quarters of enrolled participants chose a LARC method when they were counseled about effectiveness and offered their choice of method at no charge, and continuation rates were high 2 years (77% for LARC users vs 41% for non-LARC users) and 3 years (67% for LARC users vs 31% for non-LARC users) after insertion [6-8]. High rates of LARC use were also found in a cluster randomized trial of a contraceptive counseling intervention, with more women enrolled in the intervention choosing a LARC method than those in the comparison group (28% vs 17%) [9].

Despite their effectiveness at preventing pregnancy and many women's preference for them, there are a number of provider-related barriers to LARC access. For example, a recent national survey of family physicians found that less than 20% regularly inserted IUDs and 11% regularly inserted implants [3]. A similar survey of obstetricians and gynecologists found that a high percentage (more than 95%) reported providing IUDs, but only half offered the implant [4]. A review of barriers to use of LARC reported that many providers have misconceptions that restrict their willingness to provide these methods; for example, they perceive that IUDs are inappropriate for use in young and nulliparous women, are concerned that IUD use will lead to pelvic inflammatory disease, perceive that LARC methods are difficult to insert, and that smoking and hypertension are contraindications to use of implants [5]. All of these concerns are inaccurate.

Further, despite the many advantages of LARC methods, there are also special concerns that affect how the proposed performance measure should be used. There is a long history in America of coercive practices with regard to contraception, in which disadvantaged and minority women were coerced to use sterilization and/or long-acting

methods of contraception [15, 16]. Implementing a clinical performance measure for LARC methods that had a high benchmark could cause great harm by incentivizing providers to overly promote the use of LARC over other methods, and discourage use of the client-centered counseling approach recommended by CDC-OPA ([14]).

The measure is designed to address two competing demands described above, i.e., to ensure women have access to LARC methods given the many provider and systems level barriers, yet to also ensure that they are offered in a client-centered, non-coercive manner. To avoid these conflicting needs, we recommend that the performance measure focus on low (rather than high) rates of use. For example, if a reporting entity has no or very few women using LARC (e.g., less than 2%), it is likely that there are provider and/or systems barriers that are restricting her access to these methods. Another way to identify potential barriers is to compare performance across a number of reporting units, and consider whether there are barriers to LARC among those that have LARC use rates that are well below the median level of performance.

We expect that use of the proposed performance measure will change provider behavior in two main ways: (1) more providers will start screening women who come for non-family planning reasons about their pregnancy intention, and providing them contraceptive services, as needed; and (2) when providing contraceptive services, more providers will do in accordance with ACOG, AAP and CDC recommendations to inform women about the availability of LARC methods, offer client-centered education about the relative effectiveness of all methods as well as other aspects the client may want to consider, and take steps to ensure that LARC methods are readily available to the client, preferably on-site and on a same-day basis.

1. Mansour, D., P. Inki, and K. Gemzell-Danielsson, Efficacy of contraceptive methods: a review of the literature. Eur J Contracept Reprod Health Care, 2010. 15 Suppl 2: p. S19-31.
2. Trussell, J., Contraceptive failure in the United States. Contraception, 2011. 83(5): p. 397-404.
3. Nisen, M.B., et al., U.S. family physicians´ intrauterine and implantable contraception provision: results from a national survey. Contraception, 2016.
4. Luchowski, A.T., et al., Obstetrician-gynecologists and contraception: long-acting reversible contraception practices and education. Contraception, 2014. 89(6): p. 578-83.
5. Lotke, P.S., Increasing Use of Long-Acting Reversible Contraception to Decrease Unplanned Pregnancy. Obstet Gynecol Clin North Am, 2015. 42(4): p. 557-67.
6. Winner, B., et al., Effectiveness of long-acting reversible contraception. N Engl J Med., 2012. 366((21)): p. 1998-2007.
7. Diedrich, J.T., et al., Three-year continuation of reversible contraception. Am J Obstet Gynecol, 2015. 213(5): p. 662 e1-8.
8. O´Neil-Callahan, M., et al., Twenty-four-month continuation of reversible contraception. Obstet Gynecol, 2013. 122(5): p. 1083-91.
9. Harper, C.C., et al., Reductions in pregnancy rates in the USA with long-acting reversible contraception: a cluster randomised trial. Lancet, 2015.
10. Ott, M.A., G.S. Sucato, and A. Committee on, Contraception for adolescents. Pediatrics, 2014. 134(4): p. e1257-81.
11. ACOG Long-acting reversible contraception: Implants and intrauterine devices, in Practice Bulletin. 2015 (reaffirmed), American College of Obstetricians and Gynecologists: Washington, DC. p. 1-13.
12. CDC, U.S. Medical Eligibility Criteria for Contraceptive Use 2010. MMWR Recomm Rep, 2010. 59(RR-4): p. 1-86.
13. ACOG Committee on Adolescent Health Care, Adolescents and Long-Acting Reversible Contraception: Implants and Intrauterine Devices, in Committee Opinion. 2014 (reaffirmed), American College of Obstetricians and Gynecologists: Washington, DC. p. 1-7.
14. CDC, Providing Quality Family Planning Services: Recommendations of the CDC and the U.S. Office of Population Affairs. MMWR Recommendations and Reports, 2014. 63(4): p. 1-54.
15. Gold, J., Guarding Against Coercion While Ensuring Access: A Delicate Balance. Guttmacher Policy Review, 2014. 17(3).

16. Dehlendorf, C., H. Bellanca, and M. Policar, Performance measures for contraceptive care: what are we actually trying to measure? Contraception, 2015. 91(6): p. 433-7.

Numerator Statement: Women aged 15-44 years of age at risk of unintended pregnancy who were provided a long-acting reversible method of contraception (LARC), i.e., intrauterine device or implant.
Denominator Statement: All women aged 15-44 years of age who are at risk of unintended pregnancy.
Denominator Exclusions: The following categories of women are excluded from the denominator: (1) those who are infecund for non-contraceptive reasons; (2) women who had a live birth in the last 2 months of the measurement year; or (3) women were still pregnant or their pregnancy outcome was unknown at the end of the year.

Measure Type: Structure
Data Source: Administrative claims
Level of Analysis: Facility, Health Plan, Population : Regional, Population : State

IF Endorsement Maintenance – Original Endorsement Date: Most Recent Endorsement Date:

New Measure -- Preliminary Analysis

Criteria 1: Importance to Measure and Report

1a. Evidence

1a. Evidence. The evidence requirements for a *process or intermediate outcome* measure is that it is based on a systematic review (SR) and grading of the body of empirical evidence where the specific focus of the evidence matches what is being measured.

The developer provides the following evidence for this measure:

- **Systematic Review of the evidence specific to this measure?** ☒ **Yes** ☐ **No**
- **Quality, Quantity and Consistency of evidence provided?** ☒ **Yes** ☐ **No**
- **Evidence graded?** ☐ **Yes** ☒ **No**

Evidence Summary

The developer provides a diagram of the relationship between the intermediate outcome (use of long-acting, reversible contraception) and health outcomes (reduction in unintended pregnancy

- The developer describes three ungraded guidelines from CDC/OPA, ACOG and AAP recommending counseling about contraception and focusing on most effective methods.
- Two systematic reviews (SR) exist on the relationship between use of long-acting reversible methods of contraception (LARC) and unintended pregnancy.
- QQC is provided for the SR, though the quality is implied by the large number of RCTs included in the SR rather than graded for the effectiveness and impact on unintended pregnancy studies. For the evidence of effectiveness for counseling, the review did not grade the overall body of evidence but individual studies were graded by the USPSTF methodology.
- Conclusions of SRs:
    - (2011) "the most effective methods (LARC and sterilization) have a failure rate that is less than 1% per year under typical use; the moderately effective methods (shot, PPR, diaphragm) have a typical failure rate of 6-12% per year"
    - (2010) "the review broadly confirmed the hierarchy of contraceptive effectiveness in descending order as: (1) female sterilisation, long-acting hormonal contraceptives (LNG-IUS and implants); (2) Cu-IUDs with_300 mm2 surface area; (3) Cu-IUDs with5300 mm2 surface area and short-acting hormonal contraceptives (injectables, oral contraceptives, the patch and vaginal ring), and (4) barrier methods and natural methods."

- (2015) Five of the 8 studies that examined use of more effective methods after counseling/interventions found an increased rate of use in the intervention vs control/comparison conditions. Three studies found no significant impact. No studies found a decreased rate of use of more effective contraceptive methods.
- Two large studies recently published (2015) results which provide additional evidence that: (a) long-acting reversible methods of contraception (LARC) are associated with reduced risk of unintended pregnancy, and (b) that the type of counseling provided is associated with selection of LARC methods by the client.

Guidance from the Evidence Algorithm

Structure measure (Box 1) → Systematic review (Box 3) → QQC present (Box 4) →SR concludes high quality evidence

Questions for the Committee:

- *What is the relationship of this measure to patient outcomes?*
- *How strong is the evidence for this relationship?*
- *Is the evidence directly applicable to the process of care being measured?*

Preliminary rating for evidence: ☒ **High** ☐ **Moderate** ☐ **Low** ☐ **Insufficient**

1b. Gap in Care/Opportunity for Improvement and 1b. Disparities

1b. Performance Gap. The performance gap requirements include demonstrating quality problems and opportunity for improvement.

The developer provides the following statistics:

- 310 million people in the United States, 62 million (20%) are women of reproductive age, 15-44 years
- Thirty-eight million women are in need of contraceptive services because they are at risk for unintended pregnancy – that is, they are sexually active, are able to get pregnant and want to avoid or space pregnancy
- 51% of the 6.7 million pregnancies each year (3.2 million) are unintended

	Iowa Medicaid, ages 15-44	Wisconsin Medicaid, ages 15-44	PPFA, ages 15-44	OPA Title X Clients ages 15-19	OPA Title X Clients ages 20-44
Number of women/sites	44,750	118,309; 17 HMOs	838,872 363 health centers		
Provided LARC method (variation)	9.3% (8.7 to 11.1% by region, 5 to 11.4% by type of benefit)	7.2% (6.1-10.9% by HMO)	11.4% (3.5 to 20.2% by affiate, 0 to 34.7% by health center); 24 health centers less than 2%	9.3% (0 to 34.3%) 9 grantees had a LARC rate <2% for adolescents and	13.9% (0 to 35.5 %) 1 grantee had a LARC rate <2% for adults

Disparities

The developer states the following: "A special analysis of data from the National Survey of Family Growth (NSFG), 2011-

2013, was conducted to explore disparities in the use of long-acting reversible methods of contraception. This analysis suggests that there are significant differences by age (for adolescents compared to adult women) and for women who were never married (compared to women of other marital status). However, there were no significant differences by race/ethnicity, most categories of marital status, and poverty level. For more details, see the testing report."

Questions for the Committee:

- *Is there a gap in care that warrants a national performance measure?*
- *Are you aware of evidence that disparities exist in this area of healthcare?*

Preliminary rating for opportunity for improvement: ☒ **High** ☐ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 1: Importance to Measure and Report (including 1a, 1b, 1c)

1a. Evidence to Support Measure Focus

Comments:

This is an intermediate outcome measure. The evidence directly applies to the outcome being measure - which is a reduction in unintended pregnancy.

1b. Performance Gap

Comments:

Again, this measure states that there is no disparities in LARC use based on race/ethnicity and poverty level. There are disparities in LARC use based on age and women who have never been married, which reflects a need for a performance measure. Yes, data on the measure by population subgroup (age and insurance/provider status).

1c. High Priority (previously referred to as High Impact)

Comments:

N/A

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability

2a1. Reliability Specifications

2a1. Specifications requires the measure, as specified, to produce consistent (reliable) and credible (valid) results about the quality of care when implemented.

Data source(s): Administrative claims

Specifications:

- The numerator is Women aged 15-44 years of age at risk of unintended pregnancy who were provided a long-acting reversible method of contraception (LARC), i.e., intrauterine device or implant.
- The denominator is All women aged 15-44 years of age who are at risk of unintended pregnancy.
- There are three exclusions: (1) those who are infecund for non-contraceptive reasons; (2) women who had a live birth in the last 2 months of the measurement year; or (3) women were still pregnant or their pregnancy outcome was unknown at the end of the year.
- The developers tested the measure specifications based on 2014 codes, but have also included the codes needed to calculate the measure using ICD-10 and 2015 NDC codes. Both sets of codes are attached. The goal was to convert the measure to a new code set, fully consistent with the intent of the original measure.

- The measure is stratified by age to separate adolescents from adults; the developer states this is for purposes of quality improvement, not risk adjustment, because of the large potential negative impact of teen pregnancy as well as the existence of special programs targeting this population.
- The calculation algorithm is included.

Questions for the Committee:

- *Are all the data elements clearly defined? Are all appropriate codes included?*
- *Is the logic or calculation algorithm clear?*
- *Is it likely this measure can be consistently implemented?*

2a2. Reliability Testing Testing attachment

2a2. Reliability testing demonstrates if the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise enough to distinguish differences in performance across providers.

SUMMARY OF TESTING

Reliability testing level ☒ **Measure score** ☐ **Data element** ☐ **Both**

Reliability testing performed with the data source and level of analysis indicated for this measure ☒ **Yes** ☐ **No**

Method(s) of reliability testing

- Testing was done using Medicaid datasets from Iowa and Louisiana and data from PPFA. Testing was performed at the health plan level and public health region levels of analysis.
- The developers describe their signal-to-noise testing approach using the Intraclass Correlation (ICC) with a Spearman-Brown prophesy formula to estimate reliability. This is an appropriate method for estimating the reliability of the measure score. Generally, reliability estimates > 0.70 are considered acceptable.

Results of reliability testing

- The developers provide a table of results of the reliability testing at five levels (affiliate, health center, health plan, public health region, and region by benefit type), stratified by age (15-20, 21-44, and 15-44)
- The developers report that "The estimated reliabilities remain above .90 for affiliates, for 22 of 25 affiliate groupings of health centers, for benefit type, for region by benefit type, and 2 of 3 age groups at the health plan level; ICCs at these levels were moderately high, ranging from 4-8%. Of note, reliability did decline <.90 at three levels, i.e., for three of the 25 affiliate groupings of health centers, among public health region and for the age group of 15-20 for health plan. However, two of the three affiliate groupings with lower reliability had only a single health center and thus no reliability estimation was possible. The ICC for public health region was also below .01; yet due to the larger number of cases for region, reliabilities remained above .70."
- The developers note that large numbers of cases allow for high reliability of the measure.

Guidance from the Reliability Algorithm

Precise specifications (Box 1)→ empirical reliability testing (Box 2) → testing of computed measure score (Box 4)→ appropriate method (Box 5) → high certainty that scores or reliable within the reporting rates recommended by the developer (Box 6a) → high

Questions for the Committee:

- *Is the test sample adequate to generalize for widespread implementation?*
- *Do the results demonstrate sufficient reliability so that differences in performance can be identified?*

Preliminary rating for reliability: ☒ **High** ☐ **Moderate** ☐ **Low** ☐ **Insufficient**

2b. Validity

2b1. Validity: Specifications

2b1. Validity Specifications. This section should determine if the measure specifications are consistent with the evidence.

Specifications consistent with evidence in 1a. ☒ **Yes** ☐ **Somewhat** ☐ **No**

Question for the Committee:

- *Are the specifications consistent with the evidence?*

2b2. Validity testing

2b2. Validity Testing should demonstrate the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality.

SUMMARY OF TESTING

Validity testing level ☒ **Measure score** ☐**Data element testing against a gold standard** ☐ **Both**

Method of validity testing of the measure score:

- ☒ **Face validity only**
- ☐ **Empirical validity testing of the measure score**

Validity testing method:

Face validity was performed systematically by 9 experts, who were not involved in developing the measure.

Validity testing results:

The panel rated the validity of the measure on a 1-5 scale.

- The mean rating for this measure was 4.33 with a median of 4.5 (between Agree and Strongly Agree), range 3-5. There were 44.4% (n = 4) of respondents who strongly agreed, 44.4% (n = 4) of respondents who agreed, and 11.1% (n = 1) of respondents who neither agreed nor disagreed that the scores obtained from this measures, as specified, will provide an accurate reflection of quality and can be used to distinguish good and poor quality in contraceptive services.
- One respondent pointed out that "quality of the indicator will in part depend on how well 'unintended' is characterized" and another said this measure "provides a good metric for access, not necessarily quality."

Questions for the Committee:

- *Is the test sample adequate to generalize for widespread implementation?*
- *Do the results demonstrate sufficient validity so that conclusions about quality can be made?*
- *Do you agree that the score from this measure as specified is an indicator of quality?*

2b3-2b7. Threats to Validity

2b3. Exclusions:

The developer did not formally test the exclusions because "The rationale for exclusion was due to the fact that some women are not at risk of unintended pregnancy due to infecundity or pregnancy." They included the following table of exclusions:

	Number of women

		PPFA, 2014	IME, 2013	WMP, 2014
Women 15-44 years of age		950,647	49,232	132,940
Exclusions	Infecund for non-contraceptive reasons	83	169	2,025
	Had a live birth in the last 2 months of the measurement year	7	520	2,995
	Pregnant or their pregnancy outcome was unknown at the end of the measurement year	111,685	3793	9,611
Number of women 15-44 years of age, after exclusions		838,872	44,750	118,309

Questions for the Committee:

- *Are the exclusions consistent with the evidence?*
- *Are any patients or patient groups inappropriately excluded from the measure?*
- *Are there additional exclusions that should be included?*
- *Are the exclusions/exceptions of sufficient frequency and variation across providers to be needed (and outweigh the data collection burden)?*

2b4. Risk adjustment: **Risk-adjustment method** ☒ **None** ☐ **Statistical model** ☐ **Stratification**

Conceptual rationale for SDS factors included ? ☒ **Yes** ☐ **No**

SDS factors included in risk model? ☐ **Yes** ☒ **No**

Risk adjustment summary

The developer included the following statement:

> "We do not believe that risk adjustment is justified. Although there are [possible] variations in contraceptive use by socio-demographic characteristics, the reason for those patterns is based on modifiable clinical and programmatic considerations rather than differing biological responses to contraception. Although providers may see some local variations by socio-demographic characteristics, we do not believe that these differences will be maintained if contraceptive services are offered in a client-centered manner, as defined by CDC-OPA recommendations for providing quality family planning services (CDC-OPA, 2014)."

Questions for the Committee:

- *If a justification for no risk adjustment is provided, is there any evidence that contradicts the developer's rationale and analysis?*
- *Do you agree with the developer's rationale that there is no conceptual basis for adjusting this measure for SDS factors?*

2b5. Meaningful difference *(can statistically significant and clinically/practically meaningful differences in performance measure scores can be identified)*:

- The developer states that "since our data contain the entirety of the defined population, estimation of sampling error and related inferential statistics such as confidence intervals are not applicable."
- They indicate that there are "very large and meaningful differences in rates across all reporting units" since performance in PPFA affiliates ranged from 3%-20% and PPFA health centers ranged from 0-44%.
- Within the Iowa Medicaid program, the differences between LARC provision in the general Medicaid program was substantially lower than in the waiver program, i.e., from approximately 5% to 11.5%. The ranges across regions were more narrow but still notable, i.e., from approximately 5-8% on the lower end and up to 13% on the higher end of the range.
- In Wisconsin, the rates across health plans ranged from 4.8% to 12.2%.

Question for the Committee:

○ *Does this measure identify meaningful differences about quality?*

2b6. Comparability of data sources/methods: N/A

2b7. Missing Data

- The developer states "The data source for this measure is claims data. Due to the nature of claims data (i.e., for billing purposes), there is typically very little missing data; further, it is difficult to ascertain when claims data is or is not missing."

Guidance from the algorithm:

Consistent with specifications (Box 1) →potential threats to validity addressed (Box 2) →face validity systematically assessed (Box 4) → substantial agreement (Box 5) → moderate (in the absence of empirical validity testing, moderate is the highest rating possible)

Preliminary rating for validity: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 2: Scientific Acceptability of Measure Properties (including all 2a, 2b, and 2d)

2a1. & 2b1. Specifications

Comments:

All data elements are clearly defined. The measure is also stratified by age to separate adolescents from adults. Appropriate ICD-10 and 2015 NDC codes are included. Reliability testing also remains generally high and, given the large numbers of cases, provides greater reliability.

Specifications are consistent with the evidence.

2a2. Reliability Testing

Comments:

Measure score testing was used to generalize for widespread performance across providers. The results do demonstrate sufficient reliability across five different levels.

2b2. Validity Testing

Comments:

Face validity, again, was determined by 9 experts (similarly to measure 2903). The mean rating was 4.33 - which indicates strongly agree that this will indicate quality. I agree with one of the respondents who said that this measure "provides a good metric for access, not necessarily quality" - as there are many different factors that contribute to provider quality of care, though this is certainly one of them.

2b3. Exclusions Analysis

2b4. Risk Adjustment/Stratification for Outcome or Resource Use Measures

2b5. Identification of Statistically Significant & Meaningful Differences In Performance

2b6. Comparability of Performance Scores When More Than One Set of Specifications

2b7. Missing Data Analysis and Minimizing Bias

Comments:

**-There was not a formal test of exclusions.

-Risk adjustment not found to be justified.

-There was some meaningful difference between LARC provision in the general Medicaid population and the waiver program, which makes sense since the waiver program's primary focus is contraceptive care.

- Comparability of performance scores was N/A.
- Missing data would be hard to ascertain since this is claims data.**

Criterion 3. Feasibility
Maintenance measures – no change in emphasis – implementation issues may be more prominent

3. Feasibility is the extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

- This measure is based on administrative claims. Claims data are generally thought to be of minimal burden and highly feasible.
- The measure has been piloted in two state Medicaid programs and is currently being used by 13 state Medicaid programs funded by the Center for Medicaid and CHIP Services. Overall reports indicate it is feasible to calculate and use the measure.
- The developer also notes some lessons learned thus far.
- No fees or licensing requirements to use the measure.

Questions for the Committee:

- *Are the required data elements routinely generated and used during care delivery?*
- *Is the data collection strategy ready to be put into operational use?*

Preliminary rating for feasibility: ☒ **High** ☐ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 3: Feasibility

3a. Byproduct of Care Processes
3b. Electronic Sources
3c. Data Collection Strategy
Comments:
The data required are routinely generated and/or used during care delivery and also do not represent an undue burden to collect. As such, this measure would be implemented without much administrative burden. Thus, feasibility is high.

Criterion 4: Usability and Use
Maintenance measures – increased emphasis – much greater focus on measure use and usefulness, including both impact /improvement and unintended consequences

4. Usability and Use evaluate the extent to which audiences (e.g., consumers, purchasers, providers, policymakers) use or could use performance results for both accountability and performance improvement activities.

Current uses of the measure

Publicly reported? ☐ **Yes** ☒ **No**

Current use in an accountability program? ☐ **Yes** ☒ **No**
OR
Planned use in an accountability program? ☒ **Yes** ☐ **No**

- OPA's Title X program intends to make publicly available the performance of all grantees (individually and aggregated by state) within a 3-year period of time.
- Internal QI:

- CMCS Maternal and Infant Health Initiative: https://www.medicaid.gov/Medicaid-CHIP-Program-Information/By-Topics/Quality-of-Care/Maternal-and-Infant-Health-Care-Quality.html
- OPA Title X family planning program: www.familyplanningdashboard.com
- Planned Parenthood Federation of America a Clinical Quality Improvement (CQI) Department (70% of affiliates)

Accountability program details
N/A

Improvement results
N/A

Unexpected findings (positive or negative) during implementation
The developer reports that no unintended negative consequences were identified.

Potential harms
The developer identified one potential concern: the measure may lead to coercive practices in which women are not offered a free choice of methods and are pressured into using a LARC method. The developer states they do not think this will be a concern for two reasons: one, the measure is focused on ensuring access to these methods by monitoring very low rates (well below the median) and the measure is not intended to be used for benchmarking. Second, "CDC-OPA recommendations describe in detail how to provide client-centered, non-coercive contraceptive counseling, and efforts to support use of the measure should be accompanied by efforts to increase awareness of the CDC-OPA recommendations (CDC/OPA 2014). Further, OPA has funded the development of training on how to provide client-centered counseling, which is available to all providers on the OPA-supported training website (www.fpntc.org)."

Feedback:
- The Measure Applications Partnership (MAP) recommended inclusion of these measures in the CMCS core adult and child measure set, pending NQF endorsement.

***Questions for the Committee**:*
- *How can the performance results be used to further the goal of high-quality, efficient healthcare?*
- *Are there any potential unintended consequences that were not identified?*
- *Do the benefits of the measure outweigh any potential unintended consequences?*

Preliminary rating for usability and use: ☒ **High** ☐ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments
Criteria 4: Usability and Use

4a. Accountability and Transparency
4b. Improvement
4c. Unintended Consequences
Comments:
This information is not currently publicly reported, however, OPA intends to make publicly available the performance of all Title X grantees within 3 years. Again, the measure developer addresses the issue of reproductive coercion, which seems to adequately address/solve that potential issue. I rate usability and use of this measure as high.

Criterion 5: Related and Competing Measures

Related or competing measures

- 2902: Contraceptive Care – Postpartum
- 2903: Contraceptive Care – Most & Moderately Effective Methods.

Harmonization

- Measures 2902, 2903 and 2904 are from the same developer and harmonized.

Pre-meeting public and member comments

- Planned Parenthood Federation of America, the nation's leading provider of women's reproductive healthcare, supports the endorsement of the proposed measures. Contraception is an important and effective preventive service to reduce unintended pregnancy as well as improve birth spacing and family planning. PPFA provided de-identified data included in the application to demonstrate the reliability and validity of the measures as well the feasibility of using them for quality improvement. Currently, PPFA has already begun using a developmental version of these measures for quality improvement and looks forward to incorporate NQF endorsed measure into its portfolio of internal quality improvement work. National endorsement of these new performance measures on contraceptive care aligns with the April 2015 call by the Institute of Medicine for standardized metrics that include measuring contraceptive use to support reducing unintended pregnancy. Further, these will be the first nationally endorsed measures on contraceptive care, providing important tools to all providers who serve women of reproductive age.

NATIONAL QUALITY FORUM—Evidence (subcriterion 1a)

Measure Number (*if previously endorsed*)**:** Click here to enter NQF number
Measure Title: Contraceptive Care – Access to LARC
IF the measure is a component in a composite performance measure, provide the title of the Composite Measure here: Click here to enter composite measure #/ title

Date of Submission: 2/15/2016

Instructions

- *For composite performance measures:*
  - *A separate evidence form is required for each component measure unless several components were studied together.*
  - *If a component measure is submitted as an individual performance measure, attach the evidence form to the individual measure submission.*
- Respond to all questions as instructed with answers immediately following the question. All information needed to demonstrate meeting the evidence subcriterion (1a) must be in this form. An appendix of *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 10 pages (*incudes questions/instructions*; minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.

Note: The information provided in this form is intended to aid the Steering Committee and other stakeholders in understanding to what degree the evidence for this measure meets NQF's evaluation criteria.

1a. Evidence to Support the Measure Focus
The measure focus is evidence-based, demonstrated as follows:

- Health outcome: [3] a rationale supports the relationship of the health outcome to processes or structures of care. Applies to patient-reported outcomes (PRO), including health-related quality of life/functional status, symptom/symptom burden, experience with care, health-related behavior.
- Intermediate clinical outcome: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured intermediate clinical outcome leads to a desired health outcome.
- Process: [5] a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured process leads to a desired health outcome.
- Structure: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured structure leads to a desired health outcome.
- Efficiency: [6] evidence not required for the resource use component.

Notes

3. Generally, rare event outcomes do not provide adequate information for improvement or discrimination; however, serious reportable events that are compared to zero are appropriate outcomes for public reporting and quality improvement.
4. The preferred systems for grading the evidence are the U.S. Preventive Services Task Force (USPSTF) grading definitions and methods, or Grading of Recommendations, Assessment, Development and Evaluation (GRADE) guidelines.
5. Clinical care processes typically include multiple steps: assess → identify problem/potential problem → choose/plan intervention (with patient input) → provide intervention → evaluate impact on health status. If the measure focus is one step in such a multistep process, the step with the strongest evidence for the link to the desired outcome should be selected as the focus of measurement. Note: A measure focused only on collecting PROM data is not a PRO-PM.
6. Measures of efficiency combine the concepts of resource use and quality (see NQF's Measurement Framework: Evaluating Efficiency Across Episodes of Care; AQA Principles of Efficiency Measures).

1a.1.This is a measure of: (*should be consistent with type of measure entered in De.1*)
Outcome

☐ Health outcome: Click here to name the health outcome

☐Patient-reported outcome (PRO): Click here to name the PRO

PROs include HRQoL/functional status, symptom/symptom burden, experience with care, health-related behaviors

☒ Intermediate clinical outcome (*e.g., lab value*): Contraceptive use

☐ Process: Click here to name the process

☐ Structure: Click here to name the structure

☐ Other: Click here to name what is being measured

HEALTH OUTCOME/PRO PERFORMANCE MEASURE *If not a health outcome or PRO, skip to 1a.3*

1a.2. Briefly state or diagram the path between the health outcome (or PRO) and the healthcare structures, processes, interventions, or services that influence it.

Not a health outcome or PRO.

1a.2.1. State the rationale supporting the relationship between the health outcome (or PRO) to at least one healthcare structure, process, intervention, or service (*i.e., influence on outcome/PRO*).

Note: For health outcome/PRO performance measures, no further information is required; however, you may provide evidence for any of the structures, processes, interventions, or service identified above.

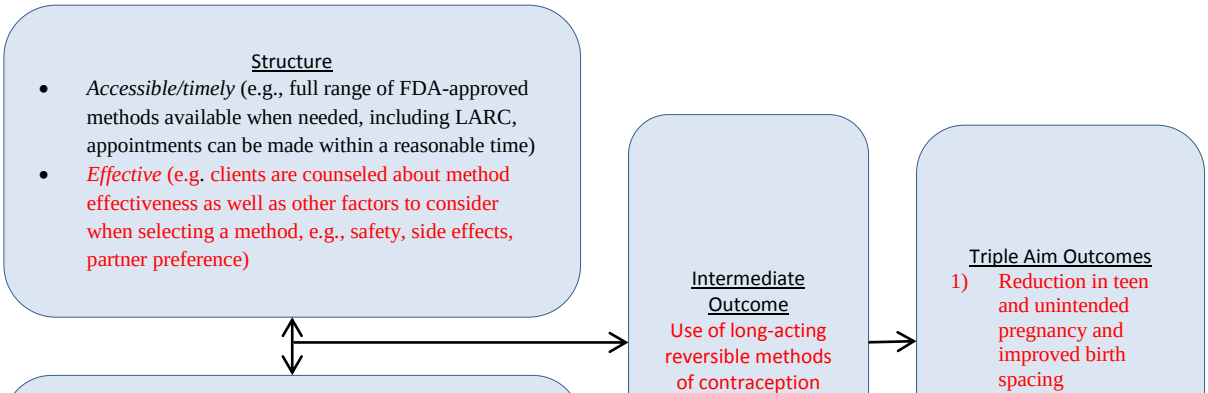
INTERMEDIATE OUTCOME, PROCESS, OR STRUCTURE PERFORMANCE MEASURE

1a.3. Briefly state or diagram the path between structure, process, intermediate outcome, and health outcomes. Include all the steps between the measure focus and the health outcome.

The diagram below illustrates the steps between the structure and process that influence the intermediate outcome, and how the intermediate outcome in turns influences the longer-term outcomes. The text highlighted in red shows the primary relationships that will be affected by use of the proposed measure: (a) increased use of the long-acting reversible methods of contraception (LARC) will influence rates of unintended pregnancy; and (b) appropriate counseling of a client can lead to increased use of the LARC methods of contraception.

The type of contraceptive method used by a woman is strongly associated with her risk of unintended pregnancy. The most effective methods (sterilization and the long-acting reversible contraceptive [LARC] methods of intrauterine devices and implants) have a failure rate that is less than 1% per year under typical use; the moderately effective methods (shot, oral pills, patch, ring, and diaphragm) have a typical failure rate of 6-12% per year; the least effective methods have a typical failure rate of 18-28%; and if no method is used then 85 of every 100 women will become pregnant in a year (Trussell 2011).

The measure is secondarily supported by evidence that the way in which contraceptive counseling is offered (e.g., increased screening of clients for reproductive intention; the provision of client-centered counseling, which includes providing information about and ready access to LARC methods; and ready access to all methods of contraception, ideally on a same-day basis) will lead to increased use of the LARC methods (i.e., the intermediate outcome).



1a.3.1. What is the source of the systematic review of the body of evidence that supports the performance measure?

☒ Clinical Practice Guideline recommendation – ***complete sections*** 1a.4***, and*** 1a.7

☐ US Preventive Services Task Force Recommendation – ***complete sections*** 1a.5 ***and*** 1a.7

☒ Other systematic review and grading of the body of evidence (*e.g., Cochrane Collaboration, AHRQ Evidence Practice Center*) – ***complete sections*** 1a.6 ***and*** 1a.7

☐ Other – ***complete section*** 1a.8

Please complete the sections indicated above for the source of evidence. You may skip the sections that do not apply.

_____

1a.4. CLINICAL PRACTICE GUIDELINE RECOMMENDATION

1a.4.1. Guideline citation (*including date*) and **URL for guideline** (*if available online*):

Clinical recommendations (from both government sources and professional organizations) are the best source of evidence about the relationship between contraceptive counseling and increased use of LARC methods (see diagram above).

CDC/OPA (2014). Providing Quality Family Planning Services (QFP): Recommendations of CDC and the US Office of Population Affairs, MMWR Recommendations and Reports, April 24, 2014. http://www.cdc.gov/mmwr/preview/mmwrhtml/rr6304a1.htm

American College of Obstetricians and Gynecologists (ACOG), Committee on Gynecologic Practice. Increasing access to contraceptive implants and intrauterine devices to reduce unintended pregnancy. Committee Opinion Number 642; October 2015.

The American Academy of Pediatrics (AAP) (2014). Contraception for Adolescents. Pediatrics, 134:e1244–e1256.

1a.4.2. Identify guideline recommendation number and/or page number and **quote verbatim, the specific guideline recommendation**.

"Providers are encouraged to present information on potential reversible methods of contraception by using a tiered approach (i.e., presenting information on the most effective methods first, before presenting information on less effective methods). This information should include an explanation that long-acting reversible contraceptive methods are safe and effective for most women, including those who have never given birth and adolescents. Information should be tailored and presented to ensure a client-centered approach. It is not appropriate to omit presenting information on

a method solely because the method is not available at the service site. If not all methods are available at the service site, it is important to have strong referral links in place to other providers to maximize opportunities for clients to obtain their preferred method that is medically appropriate."
Source: CDC/OPA (2014). *Providing Quality Family Planning Services*, page 8 and Appendix B

"For all women at risk of unintended pregnancy, obstetrician-gynecologists should provide counseling on all contraceptive options, including implants and IUDs. Long-acting reversible contraception methods require a single action of motivation for long-term use, eliminating adherence and user dependence from the effectiveness equation. These top-tier methods share the highest continuation rates of all contraceptives, which is one of the most important factors in contraceptive success."
Source: ACOG (2015), page 1.

"Contraceptive methods most commonly used by adolescents are listed below, ordered from most to least effective, starting with long-acting reversible contraception (LARC); implants and IUDs. *Pediatricians are encouraged to counsel adolescents in that order, discussing the most effective contraceptive methods first*." ACOG (2014), page e1246.

1a.4.3. Grade assigned to the quoted recommendation with definition of the grade:

Not applicable

1a.4.4. Provide all other grades and associated definitions for recommendations in the grading system. (*Note: If separate grades for the strength of the evidence, report them in section 1a.7.*)

Not applicable

1a.4.5. Citation and URL for methodology for grading recommendations (*if different from 1a.4.1*)**:**

Not applicable

1a.4.6. If guideline is evidence-based (rather than expert opinion), are the details of the quantity, quality, and consistency of the body of evidence available (e.g., evidence tables)?

☒ Yes → ***complete section*** 1a.7
☐ No → report on another systematic review of the evidence in sections 1a.6 and 1a.7; if another review does not exist, ***provide what is known from the guideline review of evidence in*** 1a.7

1a.5. UNITED STATES PREVENTIVE SERVICES TASK FORCE RECOMMENDATION
1a.5.1. Recommendation citation (*including date*) and **URL for recommendation** (*if available online*):

Not applicable

1a.5.2. Identify recommendation number and/or page number and **quote verbatim, the specific recommendation**.

Not applicable

1a.5.3. Grade assigned to the quoted recommendation with definition of the grade:

Not applicable

1a.5.4. Provide all other grades and associated definitions for recommendations in the grading system. (*Note: the grading system for the evidence should be reported in section 1a.7.*)

Not applicable

1a.5.5. Citation and URL for methodology for grading recommendations (*if different from 1a.5.1*)**:**

Not applicable

Complete section 1a.7

_____

1a.6. OTHER SYSTEMATIC REVIEW OF THE BODY OF EVIDENCE

1a.6.1. Citation (*including date*) and **URL** (*if available online*):

Two systematic literature reviews are the best source of evidence about the relationship between use of long-acting reversible methods of contraception (LARC) and unintended pregnancy (see diagram in 1a.3, above).

The first review was led by Professor James Trussell from Princeton University, which is repeated on an ongoing basis and published in a handbook entitled "Contraceptive Technology". The Trussell analyses serve as the primary source of information about contraceptive failure rates, and are cited by the World Health Organization, CDC, and leading professional associations in the U.S. and in other countries. Trussell used two sources of data when estimating contraceptive failure. The first was published research, which comprised results from clinical trials and surveys. The second source was the CDC's National Survey of Family Growth (NSFG), which was used to estimate *typical* use rates using data from a nationally representative sample of users.

- Trussell J (2011). Contraceptive efficacy. In: Hatcher RA, Trussell J, Nelson AL, Cates W, Kowal D, editors. Contraceptive technology: twentieth revised edition. New York: Ardent Media; 2011, pp. 777–861. This was subsequently summarized in: Trussell J (2011). Contraceptive failure in the United States. Contraception; 83(5):397-404.

- WHO/Department of Reproductive Health and Research & Johns Hopkins Bloomberg School of Public Health/Center for Communication Programs (2011). Family Planning: A Global Handbook for Providers. Baltimore and Geneva: CCP and WHO.

The second review was conducted by Mansour et al in 2010. They search Medline and Embase from January 1990 to February 2008 for publications reporting contraceptive failure rates.

- Mansour D, Inki P, Gemzell-Danielsson K (2010). Efficacy of contraceptive methods: A review of the literature. The European Journal of Contraception and Reproductive Health Care, 15:4-16.

1a.6.2. Citation and URL for methodology for evidence review and grading (*if different from 1a.6.1*)**:**

Evidence of contraceptive effectiveness & its impact on unintended pregnancy

- See 1a.6.1 above

Evidence of effectiveness of counseling or other interventions to affect patients' choice of method

- Zapata LB, Tregear SJ, Curtis KM, Tiller M, Pazol K, Mautone-Smith N, Gavin LE (2015). Impact of Contraceptive Counseling in Clinical Settings: A Systematic Review. Am J Prev Med. 2015 Aug;49(2 Suppl 1):S31-45.

Complete section 1a.7

_____

1a.7. FINDINGS FROM SYSTEMATIC REVIEW OF BODY OF THE EVIDENCE SUPPORTING THE MEASURE

If more than one systematic review of the evidence is identified above, you may choose to summarize the one (or more) for which the best information is available to provide a summary of the quantity, quality, and consistency of the body of evidence. Be sure to identify which review is the basis of the responses in this section and if more than one, provide a separate response for each review.

1a.7.1. What was the specific structure, treatment, intervention, service, or intermediate outcome addressed in the evidence review?

Evidence of contraceptive effectiveness & its impact on unintended pregnancy

The studies examining contraceptive efficacy and effectiveness considered the impact of use of specific contraceptive methods on risk of pregnancy (i.e., contraceptive failure). Pregnancy risk can be assessed either through life table analyses (usually through 12 months) that show the percentage of women who become pregnant, or the score on the Pearl Index. The Pearl Index is a commonly used technique for reporting the effectiveness of a birth control method in clinical trials, and estimates the number of unintended pregnancies over a period of exposure (e.g. 100 women over one year of use, or 10 women over 10 years). Contraceptive failure rates are reported for *perfect use* and *typical use*. Perfect use reflects how effective methods can be in preventing pregnancy when used consistently and correctly according to instructions. Typical use reflects how effective methods are for the average person who does not always use methods correctly or consistently. Pregnancy rates during typical use of adherence-dependent methods (such as the oral pill) generally vary widely for different groups using the same method, primarily due to differences in the propensity to use the method perfectly.

Evidence of effectiveness of counseling or other interventions to affect patients' choice of method

The systematic review underpinning the CDC-OPA recommendation on contraceptive counseling used an analytic framework that considered the impact of providing contraceptive counseling and/or education on short (e.g., client knowledge, attitudes), medium (e.g., selection of more effective methods, correct and consistent use) and long-term (unintended pregnancy) outcomes (Zapata 2015).

1a.7.2. Grade assigned for the quality of the quoted evidence with definition of the grade:

Evidence of contraceptive effectiveness & its impact on unintended pregnancy

While the quality of the studies was not graded in either the Trussell (2011) or Mansour (2010) review, it was significantly based on systematic reviews of randomized controlled trials.

Evidence of effectiveness of counseling or other interventions to affect patients' choice of method

The review did not grade the *overall* body of evidence. However, the quality of individual studies was graded in accordance with USPSTF methodologies for doing so, i.e., Level I, Level II-1, Level II-2, Level II-3, Level III.

1a.7.3. Provide all other grades and associated definitions for strength of the evidence in the grading system.

Not applicable

1a.7.4. What is the time period covered by the body of evidence? (*provide the date range, e.g., 1990-2010*). Date range

Trussell (2011): 1958-2010
Mansour (2010): January 1990 to February 2008
Zapata (2015): 1985-February 2011 with supplemental searches through 2014

QUANTITY AND QUALITY OF BODY OF EVIDENCE

1a.7.5. How many and what type of study designs are included in the body of evidence? (*e.g., 3 randomized controlled trials and 1 observational study*)

Evidence of contraceptive effectiveness & its impact on unintended pregnancy

- Trussell et al 2011: The review comprised results from clinical trials and surveys; the most recent review listed more than 350 studies, of which the majority was randomized controlled trials (Trussell 2011a). Of these, 13 trials focused on failure rates for intrauterine devices, and 16 focused on contraceptive implants.

- Mansour et al 2010: The authors identified and extracted information from 139 publications. Of the included studies, 47 assessed combined oral contraceptives (COCs), one assessed progestogen-only pills (POPs), three assessed the patch, three assessed the vaginal ring, 15 assessed implants, 16 assessed injectables, 31 assessed copper intrauterine devices (Cu-IUDs), nine assessed the levonorgestrel-releasing intrauterine system (LNGIUS), three assessed the male condom, four assessed other barrier methods, 11 assessed natural methods, and four assessed female sterilization. Overall, there were 64 publications of randomized controlled studies included in this review. A detailed description of each publication can be accessed from www.informahealthcare.com/doi/pdf/10.3109/13625180903427675.

Evidence of effectiveness of counseling or other interventions to affect patients' choice of method

- Zapata et al (2015): 22 studies (from 23 articles) met the inclusion criteria; 8 studies included use of more effective methods as an outcome. Seven of the 8 studies were randomized controlled trials, while the eighth utilized a pre-posttest study design.

1a.7.6. What is the overall quality of evidence across studies in the body of evidence? (*discuss the certainty or confidence in the estimates of effect particularly in relation to study factors such as design flaws, imprecision due to small numbers, indirectness of studies to the measure focus or target population*)

The quality of evidence is not described in either the Trussell (2011) or the Mansour (2010) publications. However, both reviews are substantially comprised of randomized controlled trials.

In Zapata et al (2011), 7 of the 8 studies were graded Level I (properly designed randomized controlled trial), and the 8th study was graded Level II-3 (evidence obtained from time series, uncontrolled trial).

ESTIMATES OF BENEFIT AND CONSISTENCY ACROSS STUDIES IN BODY OF EVIDENCE

1a.7.7. What are the estimates of benefit—magnitude and direction of effect on outcome(s) across studies in the body of evidence? (*e.g., ranges of percentages or odds ratios for improvement/ decline across studies, results of meta-analysis, and statistical significance*)

Evidence of contraceptive effectiveness & its impact on unintended pregnancy

- Trussell et al 2011: The key findings of this review are estimated failure rates for a wide range of contraceptive methods under "perfect" and "typical" use. The most recent findings – published in 2011 -- are that the most effective methods (LARC and sterilization) have a failure rate that is less than 1% per year under typical use; the moderately effective methods (shot, PPR, diaphragm) have a typical failure rate of 6-12% per year; the least effective methods have a typical failure rate of 18-28%; and not using any method at all has a failure rate of 85%.

- Mansour et al 2010: "Information was identified and extracted from 139 studies. One-year Pearl Indices reported for short-acting user-dependent hormonal methods were generally less than 2.5. Gross life-table rates for long-acting hormonal methods (implants and the levonorgestrel releasing-intrauterine system [LNG-IUS]) generally ranged between 0–0.6 per 100 at one year, but wider ranges (0.1–1.5 per 100) were observed for the copper intrauterine devices (0.1–1.4 per 100 for Cu-IUDs with surface area _300 mm2 and 0.6–1.5 per 100 for those with surface area5300 mm2). Barrier and natural methods were the least effective." The authors conclude that "the review broadly confirmed the hierarchy of contraceptive effectiveness in descending order as: (1) female sterilisation, long-acting hormonal contraceptives (LNG-IUS and implants); (2) Cu-IUDs with_300 mm2 surface area; (3) Cu-IUDs with5300 mm2 surface area and short-acting hormonal contraceptives (injectables, oral contraceptives, the patch and vaginal ring), and (4) barrier methods and natural methods."

Evidence of effectiveness of counseling or other interventions to affect patients' choice of method

- Zapata (2015): Five of the 8 studies that examined use of more effective methods found an increased rate of use in the intervention vs control/comparison conditions. Three studies found no significant impact. No studies found a decreased rate of use of more effective contraceptive methods.

1a.7.8. What harms were studied and how do they affect the net benefit (benefits over harms)?

The harms were not noted in the cited reviews. However, CDC clinical recommendations on contraceptive safety explicitly address this question. CDC's "US Medical Eligibility Criteria for Contraceptive Use" (USMEC) describe what contraceptive methods are safe for women with a range of characteristics (e.g., age, postpartum) and medical conditions (e.g., infectious or chronic diseases). The citation for the USMEC recommendations is:

> CDC (2010). US Medical Eligibility Criteria for Contraceptive Use, MMWR Recommendations and Reports, 59 (RR04):1–85. Available online at: http://www.cdc.gov/reproductivehealth/UnintendedPregnancy/USMEC.htm."

The evidence on which the USMEC recommendations are based has been summarized in the following journal supplement:

> Contraception, Volume 82, Issue 1, Pages 1-118 (July 2010). Available online at: http://www.sciencedirect.com/science/journal/00107824/82/1

UPDATE TO THE SYSTEMATIC REVIEW(S) OF THE BODY OF EVIDENCE

1a.7.9. If new studies have been conducted since the systematic review of the body of evidence, provide for each new study: 1) citation, 2) description, 3) results, 4) impact on conclusions of systematic review.

Results from two large studies have been recently published, which provide additional evidence that: (a) long-acting reversible methods of contraception (LARC) are associated with reduced risk of unintended pregnancy, and (b) that the type of counseling provided is associated with selection of LARC methods by the client. The first study is a cluster-randomized trial led by researchers at the University of California – San Francisco (Harper 2015) and the second is a prospective cohort study that is known as "Project CHOICE" (Winner 2012).

UCSF trial (Harper et al 2015)

A cluster randomized trial was conducted in 2011-2013 to assess the effects of an intervention to increase patients' access to long-acting reversible contraceptives (LARCs) on pregnancy rates. A total of 40 clinics participated: 20 clinics were randomly assigned to receive evidence-based training on providing counselling and insertion of intrauterine devices (IUDs) or progestin implants, and 20 to provide standard care. Usual costs for contraception were maintained at all sites. Women aged 18-25 years attending family planning or abortion care visits and not desiring pregnancy in the next 12 months were recruited. The primary outcome was selection of an IUD or implant at the clinic visit and secondary

outcome was pregnancy within 12 months. Generalised estimating equations for clustered data were used to measure the intervention effect on contraceptive selection, and survival analysis was used to assess pregnancy rates. Of 1500 women enrolled, more at intervention than control sites reported receiving counselling on IUDs or implants (565 [71%] of 797 vs 271 [39%] of 693, odds ratio 3·8, 95% CI 2·8-5·2) and more selected LARCs during the clinic visit (224 [28%] vs 117 [17%], 1·9, 1·3-2·8). The pregnancy rate was lower in intervention group than in the control group after family planning visits (7·9 vs 15·4 per 100 person-years), but not after abortion visits (26·5 vs 22·3 per 100 person-years). We found a significant intervention effect on pregnancy rates in women attending family planning visits (hazard ratio 0·54, 95% CI 0·34-0·85).

- Harper C, Rocca CH, Thompson KM, Morfesis J, Goodman S, Darney PD, Westhoff CL, Speidel JJ (2015). Reductions in pregnancy rates in the USA with long-acting reversible contraception: a cluster randomised trial. Lancet. Volume 386, No. 9993, p562–568, 8 August 2015

Project CHOICE (Secura et al 2014, Winner et al 2015)

The Contraceptive CHOICE Project was a prospective cohort study involving 9256 St. Louis area adolescent and adult women 14 to 45 years of age, in which women were counseled about the use of LARC methods to prevent unintended pregnancy. Participants were educated about reversible contraception, with an emphasis on the benefits of LARC methods, were provided with their choice of reversible contraception at no cost, and were followed for 2 to 3 years. Almost three-quarters of enrolled participants chose a LARC method when they were counseled about effectiveness and offered their choice of method at no charge, and continuation rates were high 2 years (77% for LARC users vs 41% for non-LARC users) and 3 years (67% for LARC users vs 31% for non-LARC users) after insertion. The contraceptive failure rate among participants using pills, patch, or ring was 4.55 per 100 participant-years, as compared with 0.27 among participants using long-acting reversible contraception (hazard ratio after adjustment for age, educational level, and history with respect to unintended pregnancy, 21.8; 95% confidence interval, 13.7 to 34.9).

- Winner B, Peipert J, Qiuhong Z, Buckel C, Madden T et al (2012). Effectiveness of Long-Acting Reversible Contraception, The New England Journal of Medicine, 366 (21): 1998-2007
- Diedrich, J.T., et al., *Three-year continuation of reversible contraception.* Am J Obstet Gynecol, 2015. **213**(5): p. 662 e1-8.
- O'Neil-Callahan, M., et al., *Twenty-four-month continuation of reversible contraception.* Obstet Gynecol, 2013. **122**(5): p. 1083-91.

1a.8 OTHER SOURCE OF EVIDENCE

If source of evidence is NOT from a clinical practice guideline, USPSTF, or systematic review, please describe the evidence on which you are basing the performance measure.

Not applicable.

1a.8.1 What process was used to identify the evidence?

Not applicable.

1a.8.2. Provide the citation and summary for each piece of evidence.

Not applicable.

1. Evidence, Performance Gap, Priority – Importance to Measure and Report

Extent to which the specific measure focus is evidence-based, important to making significant gains in healthcare quality, and improving health outcomes for a specific high-priority (high-impact) aspect of healthcare where there is variation in or overall less-than-optimal performance. ***Measures must be judged to meet all subcriteria to pass this criterion and be evaluated against the remaining criteria.***

1a. Evidence to Support the Measure Focus – See attached Evidence Submission Form
nqf_evidence_Contraceptive_Care_LARC.docx

1b. Performance Gap

Demonstration of quality problems and opportunity for improvement, i.e., data demonstrating:

- considerable variation, or overall less-than-optimal performance, in the quality of care across providers; and/or
- disparities in care across population groups.

1b.1. **Briefly explain the rationale for this measure** *(e.g., the benefits or improvements in quality envisioned by use of this measure)*
Contraception is a highly effective clinical preventive service intended to prevent teen and unintended pregnancy and space births. The type of contraceptive method used by a woman is strongly associated with her risk of unintended pregnancy [1, 2]. The most effective methods (sterilization and the long-acting reversible methods of IUDs and implants) have a failure rate that is less than 1% per year under typical use; the moderately effective methods (shot, pill, patch, ring and diaphragm) have a typical failure rate of 6-12% per year; the least effective methods have a typical failure rate of 18-28%; and 85% of women will become pregnant in a year if they do not use any contraceptive method at all [2]. The proposed performance measure is based on the fact that the long-acting reversible methods of contraception (LARC), i.e., IUDs and implants, are substantially more effective than other reversible methods at preventing unintended pregnancy, and is designed to ensure women have access to these methods.

Existing research shows that many women will choose LARC methods if given the opportunity to do so. For example, in a recent large prospective study, almost three-quarters of enrolled participants chose a LARC method when they were counseled about effectiveness and offered their choice of method at no charge, and continuation rates were high 2 years (77% for LARC users vs 41% for non-LARC users) and 3 years (67% for LARC users vs 31% for non-LARC users) after insertion [6-8]. High rates of LARC use were also found in a cluster randomized trial of a contraceptive counseling intervention, with more women enrolled in the intervention choosing a LARC method than those in the comparison group (28% vs 17%) [9].

Despite their effectiveness at preventing pregnancy and many women's preference for them, there are a number of provider-related barriers to LARC access. For example, a recent national survey of family physicians found that less than 20% regularly inserted IUDs and 11% regularly inserted implants [3]. A similar survey of obstetricians and gynecologists found that a high percentage (more than 95%) reported providing IUDs, but only half offered the implant [4]. A review of barriers to use of LARC reported that many providers have misconceptions that restrict their willingness to provide these methods; for example, they perceive that IUDs are inappropriate for use in young and nulliparous women, are concerned that IUD use will lead to pelvic inflammatory disease, perceive that LARC methods are difficult to insert, and that smoking and hypertension are contraindications to use of implants [5]. All of these concerns are inaccurate.

Further, despite the many advantages of LARC methods, there are also special concerns that affect how the proposed performance measure should be used. There is a long history in America of coercive practices with regard to contraception, in which disadvantaged and minority women were coerced to use sterilization and/or long-acting methods of contraception [15, 16]. Implementing a clinical performance measure for LARC methods that had a high benchmark could cause great harm by incentivizing providers to overly promote the use of LARC over other methods, and discourage use of the client-centered counseling approach recommended by CDC-OPA ([14]).

The measure is designed to address two competing demands described above, i.e., to ensure women have access to LARC methods given the many provider and systems level barriers, yet to also ensure that they are offered in a client-centered, non-coercive manner. To avoid these conflicting needs, we recommend that the performance measure focus on low (rather than high) rates of use. For example, if a reporting entity has no or very few women using LARC (e.g., less than 2%), it is likely that there are provider and/or systems barriers that are restricting her access to these methods. Another way to identify potential barriers is to compare performance across a number of reporting units, and consider whether there are barriers to LARC among those that have LARC use rates that are well below the median level of performance.

We expect that use of the proposed performance measure will change provider behavior in two main ways: (1) more providers will start screening women who come for non-family planning reasons about their pregnancy intention, and providing them contraceptive services, as needed; and (2) when providing contraceptive services, more providers will do in accordance with ACOG, AAP and CDC recommendations to inform women about the availability of LARC methods, offer client-centered education about the relative effectiveness of all methods as well as other aspects the client may want to consider, and take steps to ensure that LARC methods are readily available to the client, preferably on-site and on a same-day basis.

1. Mansour, D., P. Inki, and K. Gemzell-Danielsson, Efficacy of contraceptive methods: a review of the literature. Eur J Contracept Reprod Health Care, 2010. 15 Suppl 2: p. S19-31.
2. Trussell, J., Contraceptive failure in the United States. Contraception, 2011. 83(5): p. 397-404.
3. Nisen, M.B., et al., U.S. family physicians´ intrauterine and implantable contraception provision: results from a national survey. Contraception, 2016.
4. Luchowski, A.T., et al., Obstetrician-gynecologists and contraception: long-acting reversible contraception practices and education. Contraception, 2014. 89(6): p. 578-83.
5. Lotke, P.S., Increasing Use of Long-Acting Reversible Contraception to Decrease Unplanned Pregnancy. Obstet Gynecol Clin North Am, 2015. 42(4): p. 557-67.
6. Winner, B., et al., Effectiveness of long-acting reversible contraception. N Engl J Med., 2012. 366((21)): p. 1998-2007.
7. Diedrich, J.T., et al., Three-year continuation of reversible contraception. Am J Obstet Gynecol, 2015. 213(5): p. 662 e1-8.
8. O´Neil-Callahan, M., et al., Twenty-four-month continuation of reversible contraception. Obstet Gynecol, 2013. 122(5): p. 1083-91.
9. Harper, C.C., et al., Reductions in pregnancy rates in the USA with long-acting reversible contraception: a cluster randomised trial. Lancet, 2015.
10. Ott, M.A., G.S. Sucato, and A. Committee on, Contraception for adolescents. Pediatrics, 2014. 134(4): p. e1257-81.
11. ACOG Long-acting reversible contraception: Implants and intrauterine devices, in Practice Bulletin. 2015 (reaffirmed), American College of Obstetricians and Gynecologists: Washington, DC. p. 1-13.
12. CDC, U.S. Medical Eligibility Criteria for Contraceptive Use 2010. MMWR Recomm Rep, 2010. 59(RR-4): p. 1-86.
13. ACOG Committee on Adolescent Health Care, Adolescents and Long-Acting Reversible Contraception: Implants and Intrauterine Devices, in Committee Opinion. 2014 (reaffirmed), American College of Obstetricians and Gynecologists: Washington, DC. p. 1-7.
14. CDC, Providing Quality Family Planning Services: Recommendations of the CDC and the U.S. Office of Population Affairs. MMWR Recommendations and Reports, 2014. 63(4): p. 1-54.
15. Gold, J., Guarding Against Coercion While Ensuring Access: A Delicate Balance. Guttmacher Policy Review, 2014. 17(3).
16. Dehlendorf, C., H. Bellanca, and M. Policar, Performance measures for contraceptive care: what are we actually trying to measure? Contraception, 2015. 91(6): p. 433-7.

1b.2. Provide performance scores on the measure as specified (current and over time) at the specified level of analysis. *(This is required for endorsement maintenance. Include mean, std dev, min, max, interquartile range, scores by decile. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included). This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*

Performance scores for the LARC access measure are presented for four programs; two are state Medicaid programs (i.e., the Iowa Medicaid Enterprise and the Wisconsin Medicaid Program) and two are organizations that focus on the delivery of reproductive health services (i.e., the Planned Parenthood Federation of America, and the OPA's Title X family planning program).

Iowa's Medicaid Enterprise (IME) provides contraceptive services to women through its general Medicaid program and its family planning waiver program (IFPN). Services are available to Iowa residents who are US citizens or qualified immigrants. To be eligible for IFPN services, the following guidelines apply: an individual does not have insurance or the insurance does not cover family planning services; the individual is a man or woman between the ages of 12 and 54; and family income is at or below 300 percent of the federal poverty level. In addition, women whose pregnancy and delivery was covered by Medicaid also qualify automatically for family planning services. A total of 44,750 women met the inclusion criteria and were included in the analysis. The results showed that 9.3% of clients ages 15-44 were provided a LARC method of contraception; there was variation by public health region (8.7 to 11.1%) and type of benefit (i.e., general Medicaid vs IFPN family planning waiver) (5 to 11.4%). For more details, see the attached testing report.

The Wisconsin Medicaid Program (WMP) provides contraceptive services to women through its general Medicaid (BadgerCare Plus)

and family planning only services (FPOS) programs. Services are available to Wisconsin residents who are US citizens or qualified immigrants meeting income eligibility criteria (e.g., a child <18 years with household income at or below 300% FPL; an adult with income at or below 100% FPL). To be eligible for FPOS, individuals must not be covered by Medicaid for the Elderly, Blind, or Disabled or BadgerCare Plus and must be at or below 300% FPL. In December 2014, 65% of Wisconsin Medicaid members were enrolled in a health maintenance organization (HMO). A total of 118,309 eligible women who participated in one of 17 HMOs were included in the analysis. The results showed that 7.2% of clients ages 15-44 were provided a LARC method of contraception; there was variation by HMO (6.1 to 10.9%). For more details, see the attached testing report.

Planned Parenthood of America (PPFA) is comprised of 66 independently incorporated affiliates, operating approximately 700 health centers in the United States, providing reproductive health care to nearly 2.7 million patients. De-identified, encounter-level data are captured in a quality information warehouse for a subset of affiliates. Data included in this analysis covers services provided at 25 affiliates between January 1 and December 31 2014. The final dataset analyzed included 838,872 female patients aged 15-44 years, who were cared for in one of 363 health centers, in the calendar year 2014. The results showed that 11.4% of clients ages 15-44 were provided a LARC method of contraception; there was variation by affiliate (3.5 to 20.2%) and health center (0 to 34.7%); 24 health centers had LARC provision rates that were less than 2%. For more details, see the attached testing report.

The Title X Family Planning program was enacted in 1970, and is the only federal grant program dedicated solely to providing low-income individuals with comprehensive family planning and related preventive health services. The U.S. Department of Health and Human Services' Office of Population Affairs (OPA) oversees the Title X program. In 2014, grantees oversaw 4,100 family planning centers which served 4.2 million clients. Services are provided through state, county, and local health departments; community health centers; dedicated reproductive health centers; and hospital-based, school-based, faith-based, other private nonprofits. The scores were based on data reported to OPA in the annual Title X Family Planning Annual Report (FPAR), rather than claims data. The Title X data is included in this application to demonstrate that even in a program that is committed to the provision of family planning services, there is substantial room for improvement in the delivery of contraceptive services. The FPAR data has several advantages over claims data, in that it documents sterilization or LARC insertion in a year preceding the measurement year, and whether the client was seeking pregnancy. The results showed that overall 9.3% of clients ages 15-19 and 13.9% of clients ages 20-44 were provided a LARC method of contraception. There was substantial variation by grantee (e.g., from 0 to 34.3% for adolescent clients, and from 0 to 35.5 % among adult clients); 9 grantees had a LARC rate <2% for adolescents and 1 grantee had a LARC rate <2% for adults. For more details, see the attached appendix.

The primary intent of the LARC measure is to identify populations in which LARC use is noticeably low so that health programs can determine if there are barriers to access. Hence, the LARC measure is used to assess access by identifying very low rates of LARC use (e.g., less than 2%) or by calculating the median or mean and then identifying those entities (e.g., public health regions, health centers, health plans) where the rates of LARC use are well below the median or mean. The developers of the measure contend that the measure should be used only to monitor access to LARC; and that it could be harmful to set a high benchmark for this measure, because doing so may incentivize coercive practices (Dehlendorf 2015, Gold 2014).

To illustrate how the LARC scores should be interpreted, we can review the scores for the LARC measure in the Iowa Medicaid population. The use of LARC does not fall below 2% in any public health region and there do not appear to be substantial differences in LARC access across public health regions. These data suggest that there is some access to LARC in the state overall and across regions of the state. However, the LARC rate was almost twice as high in the state family planning waiver program compared to the general Medicaid program; this suggests that there it may be worth investigating potential barriers to LARC provision in the general Medicaid program.

Data from the Planned Parenthood and Title X programs show that overall LARC rates were relatively high. A few Planned Parenthood health centers and Title X grantees had LARC rates below 2%, which suggests that there may be some locations in which clients may not have adequate access to LARC methods. Due to the availability of the LARC access measure, Planned Parenthood and OPA are able to identify and follow up with these grantees to assess what barriers may exist, and determine how to overcome them so that clients are given the opportunity to obtain LARC if they choose to do so.

1b.3. If no or limited performance data on the measure as specified is reported in 1b2, then provide a summary of data from the literature that indicates opportunity for improvement or overall less than optimal performance on the specific focus of measurement.

1b.4. **Provide disparities data from the measure as specified (current and over time) by population group, e.g., by race/ethnicity, gender, age, insurance status, socioeconomic status, and/or disability.** *(This is required for endorsement maintenance. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities include.) This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*

1b.5. **If no or limited data on disparities from the measure as specified is reported in 1b4, then provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement. Include citations.**

A special analysis of data from the National Survey of Family Growth (NSFG), 2011-2013, was conducted to explore disparities in the use of long-acting reversible methods of contraception. This analysis suggests that there are significant differences by age (for adolescents compared to adult women) and for women who were never married (compared to women of other marital status). However, there were no significant differences by race/ethnicity, most categories of marital status, and poverty level. For more details, see the testing report.

1c. High Priority (previously referred to as High Impact)

The measure addresses:

- a specific national health goal/priority identified by DHHS or the National Priorities Partnership convened by NQF; OR
- a demonstrated high-priority (high-impact) aspect of healthcare (e.g., affects large numbers of patients and/or has a substantial impact for a smaller population; leading cause of morbidity/mortality; high resource use (current and/or future); severity of illness; and severity of patient/societal consequences of poor quality).

1c.1. **Demonstrated high priority aspect of healthcare**

Affects large numbers, A leading cause of morbidity/mortality, Frequently performed procedure, High resource use, Patient/societal consequences of poor quality

1c.2. **If Other:**

1c.3. **Provide epidemiologic or resource use data that demonstrates the measure addresses a high priority aspect of healthcare. List citations in 1c.4.**

Over the course of a lifetime, most individuals will make decisions related to childbearing, i.e., how to prevent or achieve pregnancy so that they can attain their desired number and spacing of children. Of the 310 million people in the United States, 62 million (20%) are women of reproductive age, 15-44 years; approximately 20% are men of that same age (US Census Bureau).

Thirty-eight million women are in need of contraceptive services because they are at risk for unintended pregnancy – that is, they are sexually active, are able to get pregnant and want to avoid or space pregnancy (Frost 2015). Contraception is highly effective at preventing unintended pregnancy, and the most effective methods of long-acting reversible contraception (i.e., intrauterine devices and implants), have a failure rate that is less than 1% (Trussell 2011). Numerous studies have documented the cost-effectiveness of contraception, with $4-8 saved for every $1 invested (Cleland 2011, Frost 2008, Thomas 2012, Trussell 2012, Trussell 2013). Yet 54% of women at risk of unintended pregnancy are either using any method or are using a least effective method of contraception (such as condoms, periodic abstinence, rhythm and withdrawal); and among those who use contraception, many do so inconsistently (CDC/NCHS special analysis, Trussell 2011, Jones 2012). The NSFG 2011-2013 showed that only about 9% of women ages 15-44 at risk of unintended pregnancy are using a LARC method (CDC/NCHS, special analysis).

Due to these patterns of contraceptive use, the rate of unintended pregnancy is high in the United States. More than one-half (51%) of the 6.7 million pregnancies each year (3.2 million) are unintended (Finer & Zolna 2014). Despite recent reductions, each year more than 58 of every 1000 women aged 15-19 years become pregnant (Curtin 2015), and 250,000 adolescents give birth (Hamilton 2015). As a result, many teen mothers will achieve less education and lower incomes while their children may experience higher rates of negative outcomes such as poorer health, lowered academic achievement, and higher rates of teen pregnancy for female children and incarceration for male children (Hoffman & Maynard 2008, Sawhill et al 2014). Taxpayers also pay a high price for the nation's high rate of teen and unintended pregnancy. For example, the cost of teen pregnancy alone has been estimated at $9.4 billion per year (Hoffman 2010). Two-thirds of births resulting from unintended pregnancies among women of all ages—more than one million births—are publicly funded; the direct medical cost of those births was estimated at $21 billion in 2010 (this figure includes costs for prenatal care, labor and delivery, post-partum care, and one year of infant care) (Sonfield 2015). Further, family planning services can improve infant health, including a reduction in the rate of preterm and low birth weight infants (Tsui 2010, Gipson 2008, Conde-Agudela 2006, Shah 2011, Zhu 1999). Approximately 1 out of every 8 pregnancies in the United States results in

preterm birth, and infant mortality rates remain high relative to other developed countries (MacDorman 2008, Martin 2015). Thus, by using contraceptive services to space births, and by offering preconception health services as part of family planning, it is possible to improve the health of infants, of women, and of men (CDC 2006).

The prevention of teen and unintended pregnancy and improved rates of birth spacing have repeatedly been identified as national priorities. Most recently, in 2015 the Institute of Medicine (IOM) recognized the importance of unintended pregnancy when they included it as one of 15 core measures that constitute the most vital signs for the nation's health and health care (IOM 2015). The United States' National Prevention Strategy (National Prevention Council) and Healthy People 2020 Objectives also include several specific objectives focused on unintended pregnancy and use of contraception.

1c.4. Citations for data demonstrating high priority provided in 1a.3

- American College of Obstetricians and Gynecologists, Committee on Gynecologic Practice. Increasing access to contraceptive implants and intrauterine devices to reduce unintended pregnancy. Committee Opinion Number 642; October 2015.
- American College of Obstetricians and Gynecologists (2011). Increasing Use of Contraceptive Implants and Intrauterine Devices to Reduce Unintended Pregnancy. ACOG Committee Opinion, Number 450, December 2009 (Reaffirmed 2011).
- American College of Obstetricians and Gynecologists (2012). Adolescents and Long-Acting Reversible Contraception: Implants and Intrauterine Devices. ACOG Committee Opinion, Number 539.
- American Academy of Pediatrics (2014). Contraception for Adolescents. Pediatrics, 134:e1244–e1256.
- CDC. Recommendations to improve preconception health and health care – United States. MMWR 2006;55(RR-06):1-23.
- CDC/OPA (2014). Providing Quality Family Planning Services: Recommendations of CDC and the US Office of Population Affairs, MMWR Recommendations and Reports, April 24, 2014, 63(RR04);1-29. Available online at: http://www.cdc.gov/reproductivehealth/UnintendedPregnancy/QFP.htm.
- Cleland K, Peipert JF, Westhoff C, Spear S, Trussell J. Family planning as a cost-saving preventive health service. N Engl J Med 2011;364(18):e37.
- Conde-Agudelo A, Rosas-Bermudez A, Kafury-Goeta AC. Birth spacing and risk of adverse perinatal outcomes: a meta-analysis. Jama. Apr 19 2006;295(15):1809-1823.
- Curtin SC, Abma JC, Kost K. (2015). 2010 pregnancy rates among U.S. women. NCHS health e-stat.
- Finer LB, Zolna MR. Shifts in intended and unintended pregnancies in the United States, 2001-2008. Am J Public Health. 2014 Feb;
- Frost JJ et.al, Contraceptive Needs and Services, 2013 Update, New York: Guttmacher Institute, 2015, Available online at: www.guttmacher.org/pubs/win/.
- Frost J, Finer L, Tapales A. The Impact of Publicly Funded Family Planning Clinic Services on Unintended Pregnancies and Government Cost Savings. Journal of Health Care for the Poor and Underserved 2008;19(3):778-796.
- Gipson JD, Koenig MA, Hindin MJ (2008). The effects of unintended pregnancy on infant, child, and parental health: a review of the literature. Stud Fam Plann. 2008 Mar;39(1):18-38. Review.
- Hamilton BE, Martin JA, Osterman MJK, et al. Births: Final data for 2014. National vital statistics reports; vol 64 no 12. Hyattsville, MD: National Center for Health Statistics. 2015.
- Hoffman S, Maynard R. Kids having kids: Economic costs and social consequences of teen pregnancy 2nd ed. Washington, DC: Urban Institute Press 2008.
- Hoffman SD, 2010. Counting It Up: The Public Costs of Teen Childbearing. [cited 2014 February 7]; Available from: http://www.thenationalcampaign.org/costs/default.aspx
- Institute of Medicine, Committee on Core Metrics for Better Health at Lower Cost (2015). Vital Signs: Core Metrics for Health and Health Care Progress. The National Academy of Sciences, Washington DC.
- Jones J, Mosher W, Daniels K. Current contraceptive use in the United States, 2006–2010, and changes in patterns of use since 1995. National health statistics reports 2012(60).
- Macdorman M, Mathews T. Recent trends in infant mortality in the United States. In: NCHS data brief, no. 9. Hyattsville, MD: US Department of Health and Human Services, CDC; 2008.
- National Prevention Council. National Prevention Strategy. Washington DC: U.S. Department of Health and Human Services, Office of the Surgeon General, ; 2011.
- Sawhill I, Karpilow Q, Venator J (2014). The impact of unintended childbearing on future generations. Center on Children and Families at Brookings, Brookings Institution, Washington, DC.
- Shah PS, Balkhair T, Ohlsson A, Beyene J, Scott F, Frick C. Intention to become pregnant and low birth weight and preterm birth: a systematic review. Matern Child Health J 2011;15(2):205-16.
- Sonfield A and Kost K (2015). Public Costs from Unintended Pregnancies and the Role of Public Insurance Programs in Paying for Pregnancy and Infant Care: Estimates for 2010, the Guttmacher Institute. Available from

http://www.guttmacher.org/pubs/public-costs-of-UP-2010.pdf.
• Thomas A. Three Strategies to Prevent Unintended Pregnancy. Journal of Policy Analysis and Management 2012;31(2):280-311.
• Trussell J. Contraceptive efficacy. In: Hatcher RA, Trussell J, Nelson AL, Cates W, Kowal D, editors. Contraceptive technology: twentieth revised edition. New York: Ardent Media; 2011, pp. 777–861.
• Trussell J. Contraceptive failure in the United States. Contraception 2011;83(5):397-404.
• Trussell J. Update on and correction to the cost-effectiveness of contraceptives in the United States. Contraception 2012;85(6):611.
• Trussell J, Henry N, Hassan F, Prezioso A, Law A, Filonenko A. Burden of unintended pregnancy in the United States: potential savings with increased use of long-acting reversible contraception. Contraception 2013;87(2):154-61.
• Tsui AO, McDonald-Mosley R, Burke AE (2010). Family planning and the burden of unintended pregnancies. Epidemiol Rev. 2010;32:152-74. Review.
• US Census Bureau. Age and Sex Composition in the United States: 2011. 2013 [cited 2013 October 17]; Available from: http://www.census.gov/population/age/data/2011comp.html
• US Department of Health and Human Services. Healthy People 2020 Objectives: Family Planning. [cited 2013 October 17]; Available from: http://www.healthypeople.gov/2020/topicsobjectives2020/overview.aspx?topicid=13
• Zhu BP, Rolfs RT, Nangle BE, Horan JM. Effect of the interval between pregnancies on perinatal outcomes. The New England journal of medicine. Feb 25 1999;340(8):589-594.

1c.5. **If a PRO-PM (e.g. HRQoL/functional status, symptom/burden, experience with care, health-related behaviors), provide evidence that the target population values the measured PRO and finds it meaningful.** *(Describe how and from whom their input was obtained.)*

Not applicable.

2. Reliability and Validity—Scientific Acceptability of Measure Properties

Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. ***Measures must be judged to meet the subcriteria for both reliability and validity to pass this criterion and be evaluated against the remaining criteria.***

2a.1. Specifications The measure is well defined and precisely specified so it can be implemented consistently within and across organizations and allows for comparability. eMeasures should be specified in the Health Quality Measures Format (HQMF) and the Quality Data Model (QDM).

De.5. **Subject/Topic Area** *(check all the areas that apply):*
Perinatal and Reproductive Health, Perinatal and Reproductive Health : Newborn, Perinatal and Reproductive Health : Perinatal, Prevention

De.6. **Cross Cutting Areas** *(check all the areas that apply):*
Access, Prevention

S.1. **Measure-specific Web Page** *(Provide a URL link to a web page specific for this measure that contains current detailed specifications including code lists, risk model details, and supplemental materials. Do not enter a URL linking to a home page or to general information.)*

S.2a. **If this is an eMeasure**, HQMF specifications must be attached. Attach the zipped output from the eMeasure authoring tool (MAT) - if the MAT was not used, contact staff. (Use the specification fields in this online form for the plain-language description of the specifications)
This is not an eMeasure **Attachment:**

S.2b. **Data Dictionary, Code Table, or Value Sets** *(and risk model codes and coefficients when applicable) must be attached. (Excel or csv file in the suggested format preferred - if not, contact staff)*
Attachment **Attachment:** Codes_2014_and_2015_LARC.xlsx

S.3. **For endorsement maintenance**, please briefly describe any changes to the measure specifications since last endorsement date and explain the reasons.

S.4. **Numerator Statement** *(Brief, narrative description of the measure focus or what is being measured about the target population, i.e., cases from the target population with the target process, condition, event, or outcome)*
IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
Women aged 15-44 years of age at risk of unintended pregnancy who were provided a long-acting reversible method of contraception (LARC), i.e., intrauterine device or implant.

S.5. **Time Period for Data** *(What is the time period in which data will be aggregated for the measure, e.g., 12 mo, 3 years, look back to August for flu vaccination? Note if there are different time periods for the numerator and denominator.)*
A twelve-month period of time is used.

S.6. **Numerator Details** *(All information required to identify and calculate the cases from the target population with the target process, condition, event, or outcome such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
IF an OUTCOME MEASURE, describe how the observed outcome is identified/counted. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
The target population is eligible women 15-44 years of age who were provided a long-acting reversible method of contraception (LARC). To identify the numerator, follow these steps:

Step 1 Define the numerator by identifying women who used a a long-acting reversible method of contraception (LARC) in the measurement year. To do this, use the codes in Table UCM-E.

Step 2 Adjust for LARC removals and re-insertions. The LARC methods can be removed at the woman's request so adjustments must be made to reflect this. Use the codes in Table UCM-F to identify women who had their IUD or implant removed at any point during the measurement year. Check to see if they had an IUD or implant reinserted on the same or a subsequent date through the end of the measurement year. If there is no code for reinsertion or provision of another most or moderately effective method, consider them as a non-user of LARC.

Step 3 Calculate the rates by dividing the number of women who used a most or moderately effective method of contraception by the number of women in the denominator. Calculate the rates separately for adolescents and adults.

S.7. **Denominator Statement** *(Brief, narrative description of the target population being measured)*
All women aged 15-44 years of age who are at risk of unintended pregnancy.

S.8. **Target Population Category** *(Check all the populations for which the measure is specified and tested if any):*
Children's Health, Maternal Health

S.9. **Denominator Details** *(All information required to identify and calculate the target population/denominator such as definitions, specific data collection items/responses , code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
The target population is women of reproductive age (i.e., ages 15–44 years). In a Medicaid population, this includes:

- Women in the general Medicaid program who were continuously enrolled during the measurement year, i.e., had no more than one gap in enrollment of up to 45 days. To determine continuous enrollment for a Medicaid enrollee for whom enrollment is verified monthly, the enrollee may not have more than a 1-month gap in coverage (i.e., an enrollee whose coverage lapses for 2 months is not considered continuously enrolled)
- All women participating in a state-sponsored family planning-specific Section 1115 waiver or in a family–planning specific state plan amendment (SPA) program, even if they were not continuously enrolled. This is because the primary intent of these waiver and/or SPA programs is to provide family planning services, including contraception.

S.10. **Denominator Exclusions** *(Brief narrative description of exclusions from the target population)*
The following categories of women are excluded from the denominator: (1) those who are infecund for non-contraceptive reasons; (2) women who had a live birth in the last 2 months of the measurement year; or (3) women were still pregnant or their pregnancy outcome was unknown at the end of the year.

S.11. **Denominator Exclusion Details** *(All information required to identify and calculate exclusions from the denominator such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
Follow the steps below to identify the denominator. The tables that are referenced are found in the attached Excel files (one file is for 2014 and the second is for 2015).

Step 1 Identify and exclude women who were infecund due to non-contraceptive reasons such as natural menopause or oophorectomy. To do this, use the codes listed in Table UCM-A.

Step 2 Identify women who were pregnant at any point in the measurement year by using the codes listed in Table UCM-B. We obtained this list of codes by reviewing the following documents, and including all pregnancy-related codes:

- CMS & NCHS (2011). ICD-9-CM Official Guidelines for Coding and Reporting, effective October 1, 2011. Available online at: http://www.cdc.gov/nchs/icd/icd9cm_addenda_guidelines.htm.
- CMS & NCHS (2016). ICD-10-CM Official Guidelines for Coding and Reporting FY 2016 Available online at: http://www.cdc.gov/nchs/icd/icd10cm.htm.

Step 3 Among women who were pregnant at any point in the measurement year, exclude those who:

- Had a live birth in the last 2 months of the measurement year because there may not have been an opportunity to provide them with contraception. A two-month period was selected because the American College of Obstetricians and Gynecologists (ACOG) recommends having a postpartum visit by 6 weeks, and an additional 2 weeks was added to allow for reasonable delays in attending the postpartum visit. To identify live births, use the codes listed in Table UCM-D. This list of codes is drawn from the

HEDIS measure of Prenatal and Postnatal care.

- Were still pregnant at the end of the year because they did not have a pregnancy outcome code indicating a non-live birth (Table UCM-C) or a live birth (Table UCM-D). Codes for non-live births were also drawn from the HEDIS measure of Prenatal and Postnatal Care.

Once the exclusions are applied, the denominator includes women who: were not pregnant at any point in the measurement year; were pregnant during the measurement year but whose pregnancy ended in the first 10 months of the measurement year, since there was adequate time to provide contraception in the postpartum period; or were pregnant during the measurement year but whose pregnancy ended in an ectopic pregnancy, stillbirth, miscarriage, or induced abortion.

S.12. **Stratification Details/Variables** *(All information required to stratify the measure results including the stratification variables, definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format with at S.2b)*

The primary stratification variable is age, so that adolescents can be examined separately from adult women. The is for purposes of quality improvement, and not risk adjustment. Teen pregnancy is worthy of a separate focus because of the largepotential negative impact on the life of the teen and her child(ren), and the existence of unique programs and contraceptive counseling approaches tailored to this population. In the data presented here, we used age groups that are consistent with Center for Medicaid and CHIP Servies (CMCS) reporting requirements, i.e., adolescents are defined as 15-20 years and adults are 21-44 years of age.

S.13. **Risk Adjustment Type** (Select type. Provide specifications for risk stratification in S.12 and for statistical model in S.14-15)

No risk adjustment or risk stratification

If other:

S.14. **Identify the statistical risk model method and variables** *(Name the statistical method - e.g., logistic regression and list all the risk factor variables. Note - risk model development and testing should be addressed with measure testing under Scientific Acceptability)*

We do not believe that risk adjustment is justified. Although there are some variations in LARC use by socio-demographic characteristics, the reason for those patterns is based on modifiable clinical and programmatic considerations rather than differing biological responses to contraception. More detailed information about variations in use of LARC methods by socio-demographic characteristics of women of reproductive age, can be found in the testing report.

S.15. **Detailed risk model specifications** *(must be in attached data dictionary/code list Excel or csv file. Also indicate if available at measure-specific URL identified in S.1.)*

Note: Risk model details (including coefficients, equations, codes with descriptors, definitions), should be provided on a separate worksheet in the suggested format in the Excel or csv file with data dictionary/code lists at S.2b.

S.15a. **Detailed risk model specifications** (*if not provided in excel or csv file at S.2b*)

S.16. **Type of score:**

Rate/proportion

If other:

S.17. **Interpretation of Score** *(Classifies interpretation of score according to whether better quality is associated with a higher score, a lower score, a score falling within a defined interval, or a passing score)*

Better quality = Score within a defined interval

S.18. **Calculation Algorithm/Measure Logic** *(Describe the calculation of the measure score as an ordered sequence of steps including identifying the target population; exclusions; cases meeting the target process, condition, event, or outcome; aggregating data; risk adjustment; etc.)*

Step 1 Identify all women aged 15-44 years of age who were enrolled in the health plan or program. In the case of general Medicaid, include women who were continuously enrolled (i.e., had no more than one gap in enrollment of up to 45 days). In the case of women enrolled in a family planning-specific expansion program (1115 waiver or state plan amendment), include all women

even if they do not meet the continuous enrollment criteria because the reason for their visit is related to pregnancy prevention.

Step 2 Define the denominator by excluding women who: (a) are infecund for non-contraceptive reasons; (b) had a live birth in the last 2 months of the measurement year; or (c) were still pregnant or their pregnancy outcome was unknown at the end of the year. Once exclusions are applied, the following groups of women will be included in the denominator: (a) those who were were not pregnant at any point in the measurement year; (b) those who had a live birth in the first 10 months of the measurement year; and (c) those who had a known miscarriage, stillbirth, ectopic pregnancy, or induced abortion during the measurement year.

Step 3 Define the numerator by using claims codes to identify women who adopted or continued use of a long-acting reversible method of contraception (LARC), i.e., IUD or implant. Adjust for LARC removals, in the manner specified above.

Step 4 Calculate the rates by dividing the number who used a long-acting reversible method of contraception (LARC) by the number of women in the denominator. Calculate the rates separately for adolescents and adults.

S.19. **Calculation Algorithm/Measure Logic Diagram URL or Attachment** *(You also may provide a diagram of the Calculation Algorithm/Measure Logic described above at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
Available in attached appendix at A.1

S.20. **Sampling** *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*
IF a PRO-PM, identify whether (and how) proxy responses are allowed.
The measure is based on data about all clients seen, not a sample.

S.21. **Survey/Patient-reported data** *(If measure is based on a survey, provide instructions for conducting the survey and guidance on minimum response rate.)*
IF a PRO-PM, specify calculation of response rates to be reported with performance measure results.
Not applicable.

S.22. **Missing data** (specify how missing data are handled, e.g., imputation, delete case.)
Required for Composites and PRO-PMs.
Not applicable.

S.23. **Data Source** *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*
If other, please describe in S.24.
Administrative claims

S.24. **Data Source or Collection Instrument** *(Identify the specific data source/data collection instrument e.g. name of database, clinical registry, collection instrument, etc.)*
IF a PRO-PM, identify the specific PROM(s); and standard methods, modes, and languages of administration.
Administrative claims data are used to calculate the measure. The data request should include an eligibility file, paid and denied claims with diagnosis codes and procedures codes (HCPCS, CPT, and ICD-9-PCS/ICD-10-PCS), as well as NDC codes.

S.25. **Data Source or Collection Instrument** *(available at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
Available in attached appendix at A.1

S.26. **Level of Analysis** *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*
Facility, Health Plan, Population : Regional, Population : State

S.27. **Care Setting** *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*

If other:

S.28. **COMPOSITE Performance Measure** - Additional Specifications *(Use this section as needed for aggregation and weighting rules, or calculation of individual performance measures if not individually endorsed.)*

2a. Reliability – See attached Measure Testing Submission Form **2b. Validity – See attached Measure Testing Submission Form** nqf_testing_Contraceptive_Care_LARC.docx

NATIONAL QUALITY FORUM—Measure Testing (subcriteria 2a2, 2b2-2b7)

Measure Number (*if previously endorsed*)**:** Click here to enter NQF number

Measure Title: Contraceptive Care – Access to LARC

Date of Submission: 2/15/2016
Type of Measure: Structure

☐ Composite – ***STOP – use composite testing form***	☐ Outcome (*including PRO-PM*)
☐ Cost/resource	☐ Process
☐ Efficiency	☒ Structure

Instructions

- Measures must be tested for all the data sources and levels of analyses that are specified. ***If there is more than one set of data specifications or more than one level of analysis, contact NQF staff*** about how to present all the testing information in one form.
- **For all measures, sections 1, 2a2, 2b2, 2b3, and 2b5 must be completed.**
- **For outcome and resource use measures**, section **2b4** also must be completed.
- If specified for **multiple data sources/sets of specificaitons** (e.g., claims and EHRs), section **2b6** also must be completed.
- Respond to all questions as instructed with answers immediately following the question. All information on testing to demonstrate meeting the subcriteria for reliability (2a2) and validity (2b2-2b6) must be in this form. An appendix for *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 20 pages (*incuding questions/instructions;* minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.
- For information on the most updated guidance on how to address sociodemographic variables and testing in this form refer to the release notes for version 6.6 of the Measure Testing Attachment.

Note: The information provided in this form is intended to aid the Steering Committee and other stakeholders in understanding to what degree the testing results for this measure meet NQF's evaluation criteria for testing.

2a2. Reliability testing [10] demonstrates the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise. For **PRO-PMs and composite performance measures**, reliability should be demonstrated for the computed performance score.

2b2. Validity testing [11] demonstrates that the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality. For **PRO-PMs and composite performance measures**, validity should be demonstrated for the computed performance score.

2b3. Exclusions are supported by the clinical evidence; otherwise, they are supported by evidence of sufficient frequency of occurrence so that results are distorted without the exclusion; [12]
AND

If patient preference (e.g., informed decision-making) is a basis for exclusion, there must be evidence that the exclusion impacts performance on the measure; in such cases, the measure must be specified so that the information about patient preference and the effect on the measure is transparent (e.g., numerator category computed separately, denominator exclusion category computed separately). [13]

2b4. For outcome measures and other measures when indicated (e.g., resource use):

- **an evidence-based risk-adjustment strategy** (e.g., risk models, risk stratification) is specified; is based on patient factors (including clinical and sociodemographic factors) that influence the measured outcome and are present at start of care; [14,15] and has demonstrated adequate discrimination and calibration

OR

- rationale/data support no risk adjustment/ stratification.

2b5. Data analysis of computed measure scores demonstrates that methods for scoring and analysis of the specified measure allow for **identification of statistically significant and practically/clinically meaningful** [16] **differences in performance**;

OR

there is evidence of overall less-than-optimal performance.

2b6. If multiple data sources/methods are specified, there is demonstration they produce comparable results.

2b7. For **eMeasures, composites, and PRO-PMs** (or other measures susceptible to missing data), analyses identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias.

Notes

10. Reliability testing applies to both the data elements and computed measure score. Examples of reliability testing for data elements include, but are not limited to: inter-rater/abstractor or intra-rater/abstractor studies; internal consistency for multi-item scales; test-retest for survey items. Reliability testing of the measure score addresses precision of measurement (e.g., signal-to-noise).

11. Validity testing applies to both the data elements and computed measure score. Validity testing of data elements typically analyzes agreement with another authoritative source of the same information. Examples of validity testing of the measure score include, but are not limited to: testing hypotheses that the measures scores indicate quality of care, e.g., measure scores are different for groups known to have differences in quality assessed by another valid quality measure or method; correlation of measure scores with another valid indicator of quality for the specific topic; or relationship to conceptually related measures (e.g., scores on process measures to scores on outcome measures). Face validity of the measure score as a quality indicator may be adequate if accomplished through a systematic and transparent process, by identified experts, and explicitly addresses whether performance scores resulting from the measure as specified can be used to distinguish good from poor quality.

12. Examples of evidence that an exclusion distorts measure results include, but are not limited to: frequency of occurrence, variability of exclusions across providers, and sensitivity analyses with and without the exclusion.

13. Patient preference is not a clinical exception to eligibility and can be influenced by provider interventions.

14. Risk factors that influence outcomes should not be specified as exclusions

15. With large enough sample sizes, small differences that are statistically significant may or may not be practically or clinically meaningful. The substantive question may be, for example, whether a statistically significant difference of one percentage point in the percentage of patients who received smoking cessation counseling (e.g., 74 percent v. 75 percent) is clinically meaningful; or whether a statistically significant difference of $25 in cost for an episode of care (e.g., $5,000 v. $5,025) is practically meaningful. Measures with overall less-than-optimal performance may not demonstrate much variability across providers.

1. DATA/SAMPLE USED FOR ALL TESTING OF THIS MEASURE

Often the same data are used for all aspects of measure testing. In an effort to eliminate duplication, the first five questions apply to all measure testing. If there are differences by aspect of testing,(e.g., reliability vs. validity) be sure to indicate the specific differences in question 1.7.

1.1. What type of data was used for testing? (*Check all the sources of data identified in the measure specifications and data used for testing the measure. Testing must be provided for all the sources of data specified and intended for measure implementation.* ***If different data sources are used for the numerator and denominator, indicate N [numerator] or D [denominator] after the checkbox.***)

Measure Specified to Use Data From: (***must be consistent with data sources entered in S.23***)	**Measure Tested with Data From:**
☐ abstracted from paper record	☐ abstracted from paper record
☒ administrative claims	☒ administrative claims
☐ clinical database/registry	☐ clinical database/registry
☐ abstracted from electronic health record	☐ abstracted from electronic health record
☐ eMeasure (HQMF) implemented in EHRs	☐ eMeasure (HQMF) implemented in EHRs
☐ other: Click here to describe	☐ other: Click here to describe

1.2. If an existing dataset was used, identify the specific dataset (*the dataset used for testing must be consistent with the measure specifications for target population and healthcare entities being measured; e.g., Medicare Part A claims, Medicaid claims, other commercial insurance, nursing home MDS, home health OASIS, clinical registry*).

Datasets from three programs were used for testing:

(1) The Planned Parenthood Federation of America (PPFA). In 2014, PPFA comprised 66 independently incorporated affiliates, operating approximately 700 health centers in the United States, and providing reproductive health care to nearly 2.7 million patients. De-identified, encounter-level data are captured in a quality information warehouse for a subset of affiliates. The final dataset analyzed included 838,872 female patients aged 15-44 years, who received services from 25 PPFA affiliates between January 1 and December 31, 2014. The measures were evaluated using all claims data among the eligible population, which included de-identified patient encounters, and identifiers for billing providers and health centers within affiliates. Affiliates cover geographic service areas that range from several counties within a state, a state population, and multiple states. Among the 25 affiliates included in the dataset, there were 363 health centers, and 4,467 unique billing providers nested among the health centers. These data cover diverse geographic regions and extremely large member populations, and thus may be considered reasonably representative of the U.S. population of women of reproductive age. Hence, OPA suggests the affiliate be considered a reasonable proxy for a U.S. state, for purposes of this application.

(2) The Iowa Medicaid enterprise (IME). The IME dataset comprised all female Medicaid clients aged 15-44 years who resided in 6 public health regions, participated in either fee-for-service care or in two health plans, and participated in either the general Medicaid program or the state's family planning waiver program. IME provides contraceptive services to women through its general Medicaid program and its family planning waiver program (IFPN). Services are available to Iowa residents who are US citizens or qualified immigrants. To be eligible for IFPN services, the following guidelines apply: an individual does not have insurance or your insurance does not cover family planning services; the individual is a man or woman between the ages of 12 and 54; family income is at or below 300 percent of the federal poverty level; and women whose pregnancy and delivery was covered by Medicaid will have family planning services covered. In 2013, Medicaid services in Iowa were provided primarily on a fee-for-service basis, although a small percentage of clients (approximately 2%) were provided care through one of two managed care organizations (MCO). Due to the small percentage of clients in Iowa who were enrolled in MCOs, we did not conduct reliability testing at this level in Iowa.

(3) The Wisconsin Medicaid Program (WMP). The WMP dataset is comprised of all female Medicaid clients aged 15-44 years who in 2014 resided in Wisconsin, had a paid Medicaid claim, and participated in either the general Medicaid program or the state's Family Planning Only Services (FPOS) program. The WMP provides contraceptive services to women through its general Medicaid program (BadgerCare Plus) and FPOS. FPOS members receive services on a fee for service basis only. Services are available to Wisconsin residents who are US citizens or qualified immigrants meeting income eligibility criteria (e.g., a child <18 years with household income at or below 300% FPL; an adult with income at or below 100% FPL). To be eligible for FPOS, individuals must not be covered by Medicaid for the Elderly, Blind, or Disabled or BadgerCare Plus and must be at or below 300% FPL. In December 2014, 65% of Wisconsin Medicaid members were enrolled in one of 18 health maintenance organizations (HMO).

1.3. What are the dates of the data used in testing? January 1 2013 – December 31 2014

Data from PPFA covered the period January 1 2014 – December 31 2014.
Data from IME covered the period January 1, 2013 – December 31, 2013.
Data from Wisconsin Medicaid covered the period January 1, 2014 – December 31, 2014.

1.4. What levels of analysis were tested? (*testing must be provided for all the levels specified and intended for measure implementation, e.g., individual clinician, hospital, health plan*)

Measure Specified to Measure Performance of: (*must be consistent with levels entered in item S.26*)	**Measure Tested at Level of:**
☐ individual clinician	☐ individual clinician
☐ group/practice	☐ group/practice
☒ hospital/facility/agency	☒ hospital/facility/agency
☒ health plan	☒ health plan
☒ other: Population/state equivalent, public health region, benefit type	☒ other: Population/state equivalent, public health region, benefit type

1.5. How many and which measured entities were included in the testing and analysis (by level of analysis and data source)? (*identify the number and descriptive characteristics of measured entities included in the analysis (e.g., size, location, type); if a sample was used, describe how entities were selected for inclusion in the sample*)

Reliability
The measure was tested at several levels, as shown in the table below.

Level	Number of measured entities	Data Source
Affiliate	25	PPFA
Health center	363	PPFA
Benefit type	2	IME
Public health region	6	IME
Health plan (Medicaid managed care/HMO)	17	WMP

Validity
A panel of experts assessed the measure's face validity.

1.6. How many and which patients were included in the testing and analysis (by level of analysis and data source)? (*identify the number and descriptive characteristics of patients included in the analysis (e.g., age, sex, race, diagnosis); if a sample was used, describe how patients were selected for inclusion in the sample*)

Level of analysis	Number of patients		
	15 - 20 years	**21 -44 years**	**15 - 44 years**
Affiliate (PPFA), n=25			
TOTAL	203,970	634,902	838,872
Range	294 - 42,698	1265 – 131,187	1701 – 173,885
Health centers within affiliate (PPFA), n=363			
TOTAL	203,970	634,902	838,872
Range	8 - 2984	31 – 11,391	48 – 13,335
Type of benefit (IME)			
General Medicaid	5254	9483	14737
Family planning waiver	6445	23568	30013
TOTAL	**11,699**	**33,051**	**44,750**
Public health region (IME)			
Region 1	3460	9588	13048
Region 2	1154	2906	4060
Region 3	1176	3175	4351
Region 4	1087	2887	3974
Region 5	1701	4359	6060
Region 6	3121	10136	13257
TOTAL	**11,699**	**33,051**	**44,750**
Health plan (WMP)			
MCO 1	4832	14043	18875
MCO 2	1838	5688	7526
MCO 3	920	2862	3782
MCO 4	1795	5681	7476
MCO 5	1231	3936	5167
MCO 6	219	725	944
MCO 7	558	1608	2166
MCO 8	352	1096	1448
MCO 9	1623	6164	7787
MCO 10	618	1683	2301
MCO 11	4898	15166	20064
MCO 12	1239	4290	5529
MCO 13	269	853	1122
MCO 14	2149	5596	7745
MCO 15	56	240	296
MCO 16	5114	18875	23989
MCO 17	559	1533	2092
TOTAL	**28,270**	**90039**	**118309**

1.7. If there are differences in the data or sample used for different aspects of testing (e.g., reliability, validity, exclusions, risk adjustment), identify how the data or sample are different for each aspect of testing reported below.

Not applicable

1.8 What were the patient-level sociodemographic (SDS) variables that were available and analyzed in the data or sample used? For example, patient-reported data (e.g., income, education, language), proxy variables when SDS data are not collected from each patient (e.g. census tract), or patient community characteristics (e.g. percent vacant housing, crime rate).

We assessed reliability of the measures after stratifying by age, i.e., adolescent versus adult. Teen pregnancy is worthy of a separate focus because of the large potential negative impact on the life of the teen and her child(ren), and the existence of unique programs and contraceptive counseling approaches tailored to this population. To define age

groups, we used the categories developed by the Center for Medicaid and CHIP Services (CMCS), i.e., individuals aged 15 through 20 years (15-20) were defined as adolescents, and individuals aged 21 through 44 years (21-44) were defined as adults.

2a2. RELIABILITY TESTING

***Note**: If accuracy/correctness (validity) of data elements was empirically tested, separate reliability testing of data elements is not required – in 2a2.1 check critical data elements; in 2a2.2 enter "see section 2b2 for validity testing of data elements"; and skip 2a2.3 and 2a2.4.*

2a2.1. What level of reliability testing was conducted? (*may be one or both levels*)

☐ **Critical data elements used in the measure** (*e.g., inter-abstractor reliability; data element reliability must address ALL critical data elements*)

☒ **Performance measure score** (e.g., *signal-to-noise analysis*)

2a2.2. For each level checked above, describe the method of reliability testing and what it tests (*describe the steps—do not just name a method; what type of error does it test; what statistical analysis was used*)

Several methods have been suggested to assess the reliability of provider-level performance measures (Adams, 2010; Scholle et al, 2008; Fung et al, 2010). These methods may focus on different facets of reliability such as consistency across time, consistency across raters or units, or variability at different levels of aggregation. The NQF has suggested a ***signal-to-noise*** approach as one way to evaluate measure reliability. According to Adams (2009), reliability can be assessed by the proportion of variance in a performance measure due to systemic differences across measured units (signal) in relation to random error (noise) within units.

When analytic units fall into a natural hierarchy (e.g. clients nested within health centers nested within health plan organizations), one can estimate multilevel variance components using hierarchical generalized linear modeling (HGLM) (Raudenbush and Bryk, 2002; Woltman et al, 2012). In this approach the within-provider regression coefficients are allowed to vary across providers as random effects. The covariance parameter for the random effect estimates the true between-provider variance after accounting for within-provider variance. HGLM methods are robust and well-developed for continuous outcomes, and have more recently been applied to binary outcomes (Ridout, 1999; Molenberghs et al, 2007).

In the present analyses, multi-level mixed models were fit to each dataset using a hierarchical SAS 9.3 GLIMMIX procedure with a log link function. Parameters were estimated by pseudo-maximum-likelihood using the Laplace method (Ene et al, 2012). Modeling proceeded in a top-down manner starting from the largest unit of aggregation; the variance component (random coefficient) was always estimated for the top level.

Reliability was then calculated as a function of the intraclass correlation (ICC) and the median number cases per unit, using the Spearman-Brown prophecy (Eijkenaar et al, 2013). ICCs are derived using the estimated variance component for the level of interest divided by the total variance (Wu et al 2012; He et al 2014). ICCs conceptually represent the proportion of total variation accounted for by the between-provider level, and thus follows the signal-to-noise framework suggested by NQF.

The HGLM method of estimation assumes a normally distributed error component; some authors have noted that ICCs on the logit scale can be inflated under certain circumstances when population rates are near the extremes (Wu et al, 2012). To provide more conservative estimation, medians were used in the Spearman-Brown reliability formula; the use of means would tend to bias estimates upward due to one or two atypically large provider units.

Structure of the Data

PPFA dataset. PPFA affiliates offer services within health centers. Inside each health center a group of billing providers offer care to clients. Modeling began at the topmost affiliate level (n=25), where all clients were aggregated within affiliate for the calculation of rate of LARC contraceptive use. The next level of analysis was performed within each of the 25 affiliates to examine health center rates (n=363 across all affiliates). This provided a basic 2-level structure of clients aggregated within each hierarchical unit. The top-down modeling approach enabled us to ignore small sample size problems and attribution error among individual billing providers; it also allowed us to explore the lowest level of 'granularity' for distinguishing performance among health centers of smaller size.

Iowa Medicaid Enterprise dataset. For IME data, modeling similarly proceeded from the top down starting with public health region (n=6). Unlike the PPFA data, IME data could not be examined by health facility. Instead the analysts were interested in reporting on public health region and benefit type (family planning waiver or general Medicaid benefit). Since the benefit type categories exist across regions, there is no nesting of units. Therefore, in Iowa the six regions were simply crossed with the type of benefit (n=12). Both of these crossed analyses were thought to provide useful and potentially actionable information about the interplay of regional and administrative influences on service delivery.

Wisconsin Medicaid dataset. For WMP data, modeling similarly proceeded from the top down starting with managed care organization (data from 17 of 18 HMOs was available).

2a2.3. For each level of testing checked above, what were the statistical results from reliability testing? (e.*g., percent agreement and kappa for the critical data elements; distribution of reliability statistics from a signal-to-noise analysis*)

The table below shows summary results of the reliability analyses at five levels (i.e., affiliate, health center, health plan, public health region and region by benefit type), stratified by three age categories (i.e., 15-20, 21-44, and 15-44). More detailed information about the analyses at each level can be found in Tables 1-4 (appended at the end of the form).

Level	Age group	Results		
		Median N	ICC	Reliability
Affiliate (PPFA)	15-20	4839	.0673	.9971
	21-44	11648	.0675	.9988
	15-44	16590	.0617	.9991
Health centers (estimated within each affiliate) (PPFA)	15-20	366	.0649 (median)	.6688 - .9949
	21-44	1016	.0401 (median)	.7775 - .9994
	15-44	1379	.0488 (median)	.8329 - .9994
Public health region (IME)	15-20	1438	.0055	.8887
	21-44	3767	.0017	.8666
	15-44	5205	.0022	.9197
Benefit type (IME)	15-20	5850	.0682	.9977
	21-44	16526	.0537	.9989
	15-44	22375	.0585	.9993
Region by benefit type (IME)	15-20	716	.0716	.9822
	21-44	2325	.0512	.9921
	15-44	2954	.0574	.9945
Health plan (WMP)	15-20	1231	.0043	.8414
	21-44	3936	.0082	.9702
	15-44	5167	.0067	.9721

For each level, the overall reliability was estimated using the medians as previously mentioned. ICCs, an indicator of the proportion of variance explained by the groupings, are also shown. Similar studies of hierarchical binary outcomes estimate ICCs in a typical range of .02 - .18 (Fung et al, 2010). The moderate ICCs found in our analyses, combined with the large volume of patients at most levels, tend to generate high reliability estimates. Using the 'floor' of reliability, we also calculate the minimum number of cases required to achieve acceptable reliability thresholds for each level.

The estimated reliabilities remain above .90 for affiliates, for 22 of 25 affiliate groupings of health centers, for benefit type, for region by benefit type, and 2 of 3 age groups at the health plan level; ICCs at these levels were moderately high, ranging from 4-8%. Of note, reliability did decline <.90 at three levels, i.e., for three of the 25 affiliate groupings of health centers, among public health region and for the age group of 15-20 for health plan. However, two of the three affiliate groupings with lower reliability had only a single health center and thus no reliability estimation was possible. The ICC for public health region was also below .01; yet due to the larger number of cases for region, reliabilities remained above .70.

It is commonly advised that reliability should be $\geq$ .90 for making decisions, and $\geq$ .70 for general reporting/monitoring (Eijkenaar, 2013; Adams, 2010). The Spearman-Brown prophecy allows one to test different values for ICC and patient volume per unit in order to predict expected reliability. Using an ICC value near the 20th percentile as a conservative expected correlation within units, we can compute the minimum recommended case load per level for each threshold of reliability. For example, for within-affiliate reporting of health centers, we used a conservative expected floor of .02 ICC to recommend that health centers have at least 115 patient cases for reporting rates to maintain >.70 reliability, and 450 cases to maintain >.90 reliability. The median ICC from actual data was nearly 3 times our conservative floor value (and most health centers exceeded this minimum number of cases) thus our reported reliabilities were considerably higher.

2a2.4 What is your interpretation of the results in terms of demonstrating reliability? (i.*e., what do the results mean and what are the norms for the test conducted?*)

Despite the challenges of recoding claims data to obtain contraceptive rates, having large and diverse datasets available made a positive impact on reliability. At the affiliate level, at some health centers, for benefit type, for region by benefit type, and for 2 of 3 age groups at the health plan level, we found reliabilities well above the commonly accepted .90 reliability threshold for reporting and decision-making. Of those that were below .90, only a few three were below .70, and two of those were due to having only one health center inside the affiliate.

High reliability was largely driven by two factors. First, the data exhibited adequate variation in the rates of LARC use at both the affiliate and lower levels. Second, the number of patients per unit at the affiliate level was mostly in the thousands, and at the lower levels, usually exceeded several hundred. For the IME data, the rates were much more uniform by region resulting in lower ICCs, but the volume of clients still enabled adequate reliability for distinguishing performance. When region was crossed by type of benefit the contraceptive rates were more variable among the units, so even given the smaller size of these analytic units the estimated reliabilities were higher.

In performing this analysis, we attempted to provide a conservative estimate of reliability wherever possible. Using medians rather than means, and presenting the 'floor' of reliability that may be observed for the smallest units, we bracket the results with worst-case scenarios. We further utilized a conservative value of ICC to recommend minimum patients per unit to maintain the .70 and .90 levels of reliability. In future years, analyses could examine the actual ICCs in order to make appropriate determinations about cases per unit. Yet even with these conservative methods, the 2014 data at the affiliate (state) and lower levels appears to provide sufficient reliability for reporting contraceptive rates.

2b2. VALIDITY TESTING

2b2.1. What level of validity testing was conducted? (*may be one or both levels*)

☐ **Critical data elements** (*data element validity must address ALL critical data elements*)

☒ **Performance measure score**

☐ **Empirical validity testing**

☒ **Systematic assessment of face validity of performance measure score as an indicator** of quality or resource use (*i.e., is an accurate reflection of performance on quality or resource use and can distinguish good from poor performance*)

2b2.2. For each level of testing checked above, describe the method of validity testing and what it tests (*describe the steps—do not just name a method; what was tested, e.g., accuracy of data elements compared to authoritative source, relationship to another measure as expected; what statistical analysis was used)*

We used a systematic process to assess the face validity of the performance measure, i.e., whether the corresponding measure scores correctly reflect the quality of care provided and adequately identify differences in quality. Nine experts with the following characteristics were identified: (1) expertise in the delivery of contraceptive services, as evidenced by employment in a clinical or managerial capacity for at least 3 years during which they delivered contraceptive services in a clinical setting (i.e., public and private family planning and primary care providers, or health administrators); and (2) expertise in the use of performance measures, as evidenced by participation in at least one effort to collect and use performance measurement data for the purpose of improving clinical services in the setting(s) in which they work. Below is the final list of experts who participated in the assessment:

1. Carol Brady, MA, Project Director, Florida Association of Healthy Start Coalitions, Inc.
2. Anne Burke, MD, Associate Professor, School of Medicine, Johns Hopkins Bayview Medical Center
   Vanessa Dalton, MD, MPH, Associate Professor, Director, Program on Women's Health Care Effectiveness Research, University of Michigan
3. Anne Dunlop, MD, MPH, Program Director, Preventive Medicine Division, Emory University School of Medicine
4. Daryn Eikner, MS, Vice President of Health Care Delivery, National Family Planning & Reproductive Health Association
5. Jan Engstrom, PhD, RN, CNM, WHNP-BC, Professor & Acting Chairperson, Department of Women, Children and Family Nursing, College of Nursing, Armour Academic Center
6. Mark Hathaway, MD, MPH, Senior Technical Advisor, Jhpiego – Johns Hopkins University
7. Michael Policar, MD, MPH, Clinical Professor of Obstetrics, Gynecology, and Reproductive Sciences, UCSF School of Medicine
8. Linda Wheal, Maternal Health Program Manager, Bureau of Quality Management, Illinois Department of Healthcare and Family Services

We contacted the selected experts to confirm consent to participate via email. Each expert panelist was sent a disclosure form to report any relevant financial or other competing interests; disclosures were compiled with brief biographies and shared with all panelists. Upon receipt of the disclosure form we sent the participant information about the measure specifications and other background information about the measure. Participants then participated in a webinar designed to provide important background information about the measure, how it is computed, the NQF endorsement process, and how the face validity assessment will be used in the application package that will be submitted to NQF. After reviewing the measure specifications and participating in the webinar the participants completed a survey (anonymous) that asked the following question about the measure:

The scores obtained from the measure as specified will provide an accurate reflection of quality and can be used to distinguish good and poor quality in contraceptive services: 1= Strongly Disagree

3=Neither Agree nor Disagree
5= Strongly Agree

ICD-10 Conversion:

We tested the measure specifications based on 2014 codes, but have also included the codes needed to calculate the measure using ICD-10 and 2015 NDC codes. Both sets of codes are attached. Our goal was to convert the measure to a new code set, fully consistent with the intent of the original measure. A description of how we converted from ICD-9 to ICD-10 is provided below, for each table listed in the measure specifications.

- **Sterilization for non-contraceptive reasons (Table UCM-A)**

We identified the 2015 ICD-10 codes for this table by using ICD-10 online conversion tools and confirming codes in the ICD-10-CM Expert for Physicians complete official code set, as well as with a clinical expert. These were confirmed with a clinical expert, Denise Wheeler, MS, Family Planning Director at the Iowa Department of Public Health.

- **Pregnancy codes (Table UCM-B)**

We identified the 2015 ICD-10 codes for this table by searching the NCHS/CMS publication, "ICD-10-CM Official Guidelines for Coding and Reporting, FY 2015". Pregnancy-related codes were found in "Chapter 15: Pregnancy, Childbirth and the Puerperium (O00-O9A)", and also Z codes for "outcome of delivery".

- **Known miscarriage, ectopic pregnancy, stillbirth, or induced abortion (Table UCM-C)**

These codes were identified by copying the Non-live Births Value Set from NCQA's Prenatal & Postpartum Care (PPC) measure (NQF#1517), as well as non-live birth codes in "Chapter 15: Pregnancy, Childbirth and Purperium (O00-O9A)". In the PPC measure, these codes are used to identify live births.

- **Delivery resulting in a live birth (Table UCM -D)**

These codes were identified by copying the Deliveries Value Set from NCQA's Prenatal & Postpartum Care (PPC) measure (NQF#1517), excluding extraction of products of conception retained and ectopic. In the PPC measure, these codes are used to identify live births.

- **Contraceptive codes (Tables UCM E, F and G)**

We used ICD-10 online conversion tools and confirming codes in the ICD-10-CM Expert for Physicians complete official code set. They were cross-checked against a ICD-10 conversion chart for family planning services that was prepared by Dr Michael Policar, from the University of California-San Francisco, and confirmed with a clinical expert, Denise Wheeler, MS, Family Planning Director at the Iowa Department of Public Health. NDC codes for 2015 were updated by using the codes for contraception contained in the HEDIS specifications for Chlamydia screening.

2b2.3. What were the statistical results from validity testing? (*e.g., correlation; t-test*)

The mean rating for this measure was 4.33 with a median of 4.5 (between Agree and Strongly Agree), range 3-5. There were 44.4% (n = 4) of respondents who strongly agreed, 44.4% (n = 4) of respondents who agreed, and 11.1% (n = 1) of respondents who neither agreed nor disagreed that the scores obtained from this measures, as specified, will provide an accurate reflection of quality and can be used to distinguish good and poor quality in contraceptive services. One respondent replied that he or she thinks that "the proposed measures are valid measures of quality contraceptive care for healthy women" and one responded he or she "feels STRONGLY that the adoption of these measures will promote providers' and practices' attention to reproductive planning and contraceptive care as part and parcel of women's primary health care." One respondent strongly agrees "that the measure has excellent face validity as currently specified." He or she also responded, "However, in the future, we would suggest considering the use of a look-back period using claims data to identify previous use of long-acting contraceptives." One respondent pointed out that "quality of the indicator will in part depend on how well 'unintended' is characterized." Finally, one respondent said this measure "provides a good metric for access, not necessarily quality."

2b2.4. What is your interpretation of the results in terms of demonstrating validity? (i.*e., what do the results mean and what are the norms for the test conducted?*)

We think that the responses to the face validity assessment indicate that the measure will provide an accurate reflection of quality and can be used to distinguish good and poor quality in contraceptive services.

_____

2b3. EXCLUSIONS ANALYSIS

NA ☐ no exclusions — *skip to section 2b4*

2b3.1. Describe the method of testing exclusions and what it tests (*describe the steps—do not just name a method; what was tested, e.g., whether exclusions affect overall performance scores; what statistical analysis was used*)

Exclusions were not formally tested. The rationale for exclusion was due to the fact that some women are not at risk of unintended pregnancy due to infecundity or pregnancy.

2b3.2. What were the statistical results from testing exclusions? (*include overall number and percentage of individuals excluded, frequency distribution of exclusions across measured entities, and impact on performance measure scores*)

The table below shows the number of women excluded in each of the two datasets, presented by the reason for exclusion.

		Number of women		
		PPFA, 2014	IME, 2013	WMP, 2014
Women 15-44 years of age		950,647	49,232	132,940
Exclusions	Infecund for non-contraceptive reasons	83	169	2,025
	Had a live birth in the last 2 months of the measurement year	7	520	2,995
	Pregnant or their pregnancy outcome was unknown at the end of the measurement year	111,685	3793	9,611
Number of women 15-44 years of age, after exclusions		838,872	44,750	118,309

2b3.3. What is your interpretation of the results in terms of demonstrating that exclusions are needed to prevent unfair distortion of performance results? (*i.e., the value outweighs the burden of increased data collection and analysis. Note: **If patient preference is an exclusion**, the measure must be specified so that the effect on the performance score is transparent, e.g., scores with and without exclusion*)

When combined, the total number of exclusions in each of the two data sets comprised 11.8% (PPFA), 9.1% (IME), and 11% (WMP) of all women 15-44 years of age, although the relative contribution of each type of exclusion varied by data set (e.g., live births in the last 2 months of the year were a relatively larger population in IME dataset than the PPFA dataset). These differences are likely explained by the fact that the emphasis of each program is slightly different, with the PPFA program more heavily focused on delivery of reproductive health care while the IME and WMP programs offer a wider range of primary, acute and curative care services. The number of women excluded will have a noticeable impact on the rates, and will be important to reassure providers that the measure is as 'fair' in terms of identifying the population at risk as claims data will allow it to be. For these reasons, we believe that the burden of applying the exclusion criteria is outweighed by the benefits of doing so.

_____

2b4. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES

If not an intermediate or health outcome, or PRO-PM, or resource use measure, skip to section 2b5.

2b4.1. What method of controlling for differences in case mix is used?

☒ **No risk adjustment or stratification**
☐ **Statistical risk model with** Click here to enter number of factors **risk factors**
☐ **Stratification by** Click here to enter number of categories **risk categories**
☐ **Other,** Click here to enter description

2b4.2. If an outcome or resource use measure is not risk adjusted or stratified, provide rationale and analyses to demonstrate that controlling for differences in patient characteristics (case mix) is not needed to achieve fair comparisons across measured entities.

We do not believe that risk adjustment is justified. Although there are [possible] variations in contraceptive use by socio-demographic characteristics, the reason for those patterns is based on modifiable clinical and programmatic considerations rather than differing biological responses to contraception. Although providers may see some local variations by socio-demographic characteristics, we do not believe that these differences will be maintained if contraceptive services are offered in a client-centered manner, as defined by CDC-OPA recommendations for providing quality family planning services (CDC-OPA, 2014).

A special analysis of data from the National Survey of Family Growth (NSFG), 2011-2013, was conducted to explore disparities in the use of long-acting reversible methods of contraception (see table below). This analysis suggests that there are significant differences by age (for adolescents compared to adult women) and for women who were never married (compared to women of other marital status). However, there were no significant differences by race/ethnicity, most categories of marital status, and poverty level.

Percentage of women 15-44 years of age at risk of unintended pregnancy* that used a long-acting reversible method of contraception (LARC), National Survey of Family Growth, 2011-2013

	Frequency	Weighted Frequency	Percent	95% Confidence Limits for Row Percent
Age				
15-19	15	128,000	3.21	0.67 - 5.75
20-29	243	2,038,000	12.36	10.03 - 14.69
30-44	193	2,340,000	9.06	7.12 - 11.01
Race/ethnicity				
Hispanic	140	1,060,000	11.62	8.31 - 14.93
NH White, Single race	204	2,699,000	10.05	8.01 - 12.08
NH Black, Single race	80	414,000	6.40	4.63 - 8.18
Marital status				
Married	177	2,331,000	12.13	9.52 - 14.73
Cohab	92	851,000	11.86	8.81 - 14.91
Wid/div/sep	56	529,000	10.35	4.98 - 15.72
Never married	126	796,000	5.38	4.03 - 6.73
Federal poverty level				
<100	166	1,310,000	10.29	8.01 - 12.56
100-199	107	1,035,000	10.42	7.43 - 13.42
200-399	112	1,265,000	9.37	6.75 - 11.99
400-499	22	293,000	7.91	4.23 - 11.60
500+	44	604,000	9.37	5.86 - 12.88

* Women are considered to be at risk of unintended pregnancy if they are not pregnant, not seeking pregnancy, are fecund, and have ever had sex.

2b4.3. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors (clinical factors or sociodemographic factors) used in the statistical risk model or for stratification by risk (*e.g., potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of p<0.10; correlation of x or higher; patient factors should be present at the start of care*)

We recommended stratifying the client population by age so that rates for adolescents can be tracked separately from those for adult women. We propose this stratification for purposes of QI but not as a method of risk-adjustment. Teen pregnancy is worthy of a separate focus because of the large potential negative impact on the life of the teen and her child(ren), and the existence of unique programs and contraceptive counseling approaches tailored to this population.

2b4.4a. What were the statistical results of the analyses used to select risk factors?

Not applicable.

2b4.4b. Describe the analyses and interpretation resulting in the decision to select SDS factors (e.g. prevalence of the factor across measured entities, empirical association with the outcome, contribution of unique variation in the outcome, assessment of between-unit effects and within-unit effects)

Not applicable.

2b4.5. Describe the method of testing/analysis used to develop and validate the adequacy of the statistical model or stratification approach (*describe the steps—do not just name a method; what statistical analysis was used*)

Provide the statistical results from testing the approach to controlling for differences in patient characteristics (case mix) below.
If stratified, skip to 2b4.9

2b4.6. Statistical Risk Model Discrimination Statistics (*e.g., c-statistic, R-squared*)**:**

Not applicable.

2b4.7. Statistical Risk Model Calibration Statistics (*e.g., Hosmer-Lemeshow statistic*):

Not applicable.

2b4.8. Statistical Risk Model Calibration – Risk decile plots or calibration curves:

Not applicable.

2b4.9. Results of Risk Stratification Analysis:

Not applicable.

2b4.10. What is your interpretation of the results in terms of demonstrating adequacy of controlling for differences in patient characteristics (case mix)? (i.*e., what do the results mean and what are the norms for the test conducted*)

Not applicable.

2b4.11. Optional Additional Testing for Risk Adjustment (*not required, but would provide additional support of adequacy of risk model, e.g., testing of risk model in another data set; sensitivity analysis for missing data; other methods that were assessed*)

Not applicable.

_____

2b5. IDENTIFICATION OF STATISTICALLY SIGNIFICANT & MEANINGFUL DIFFERENCES IN PERFORMANCE

2b5.1. Describe the method for determining if statistically significant and clinically/practically meaningful differences in performance measure scores among the measured entities can be identified (*describe the steps—do not just name a method; what statistical analysis was used? Do not just repeat the information provided related to performance gap in 1b)*

Due to the fact that our dataset represents a census of all claims available, rates are assumed to reflect 'true' rates by unit for the data year. Non-sampling error (such as coding or measurement error) is not estimable given our limited access to the claims data and processes. Thus we do not present any confidence intervals for inferential testing results. These assumed-true differences in rates must therefore be evaluated based on practical or clinically meaningful impact.

2b5.2. What were the statistical results from testing the ability to identify statistically significant and/or clinically/practically meaningful differences in performance measure scores across measured entities? (e.g., *number and percentage of entities with scores that were statistically significantly different from mean or some benchmark, different from expected; how was meaningful difference defined*)

The table below summarizes rates at each level. As noted above, since our data contain the entirety of the defined population, estimation of sampling error and related inferential statistics such as confidence intervals are not applicable. More detailed information about the variation of rates by unit within each level can be found in Tables 1-4, which are appended at the end of this document.

Level	Age group	Rate (use of LARC methods)
Affiliate (PPFA), n=25 Mean (range)	15-20	.099 (.029-.210)
	21-44	.119 (.028-.199)
	15-44	.114 (.035-.202)
Health center (PPFA), n=363 Mean (range)	15-20	.102 (.000-.388)
	21-44	.114 (.000-.312)
	15-44	.110 (.000-.347)
Public health region (IME) Mean (range)	15-20	.085 (.074-.104)
	21-44	.096 (.087-.113)
	15-44	.093 (.087-.111)
Benefit type (IME) Mean (range)	15-20	.085 (.047-.116)
	21-44	.096 (.051-.114)
	15-44	.093 (.050-.114)
PH Region by benefit type (IME) Mean (range)	15-20	.085 (.034-.139)
	21-44	.096 (.048-.129)
	15-44	.093 (.048-.131)
Health plan/HMO (WMP) Mean (range)	15-20	.057 (.048-.075)
	21-44	.077 (.058-.122)
	15-44	.072 (.061-.109)

2b5.3. What is your interpretation of the results in terms of demonstrating the ability to identify statistically significant and/or clinically/practically meaningful differences in performance across measured entities? (i.*e., what do the results mean in terms of statistical and meaningful differences?*)

There are very large and meaningful differences in rates across all reporting units. For example, the provision of LARC across affiliates ranged from approximately 3% to 20% within the 15-44 year age group. Among health centers, the range was 0% to almost 40% within the 15-44 year age group; four health centers had 0% LARC use and 24 had LARC use that was less than 2%. Within the IME program, the differences between LARC provision in the general Medicaid program was substantially lower than in the waiver program, i.e., from approximately 5% to 11.5%. The ranges across regions were more narrow but still notable, i.e., from approximately 5-8% on the lower end and up to 13% on the higher end of the range. In Wisconsin, the rates across health plans ranged from 4.8% to 12.2%.

These differences suggest that it will be possible to use these measures to identify meaningful differences in performance across measured entities. For example, the PPFA health centers with LARC use that is below 2% could be assessed to identify avoidable barriers to LARC access, and steps could be taken to remove those barriers. In Iowa, it may be useful to explore why LARC provision is so much lower in the general Medicaid program than in the family planning waiver program; and regions that are well below the median should be similarly assessed to see if steps can be taken to improve clients' access to LARC. Similarly, in Wisconsin, health plans with LARC provision rates that are below the mean could be assessed to determine if there are barriers that could be removed.

2b6. COMPARABILITY OF PERFORMANCE SCORES WHEN MORE THAN ONE SET OF SPECIFICATIONS
***If only one set of specifications, this section can be skipped*.**

Not applicable.

Note: *This item is directed to measures that are risk-adjusted (with or without SDS factors)* ***OR*** *to measures with more than one set of specifications/instructions (e.g., one set of specifications for how to identify and compute the measure from medical record abstraction and a different set of specifications for claims or eMeasures). It does not apply to measures that use more than one source of data in one set of specifications/instructions (e.g., claims data to identify the denominator and medical record abstraction for the numerator).* ***Comparability is not required when comparing performance scores with and without SDS factors in the risk adjustment model. However, if comparability is not demonstrated for measures with more than one set of specifications/instructions, the different specifications (e.g., for medical records vs. claims) should be submitted as separate measures.***

2b6.1. Describe the method of testing conducted to compare performance scores for the same entities across the different data sources/specifications (*describe the steps—do not just name a method; what statistical analysis was used*)

Not applicable.

2b6.2. What were the statistical results from testing comparability of performance scores for the same entities when using different data sources/specifications? (*e.g., correlation, rank order*)

Not applicable.

2b6.3. What is your interpretation of the results in terms of the differences in performance measure scores for the same entities across the different data sources/specifications? (i.*e., what do the results mean and what are the norms for the test conducted*)

Not applicable.

2b7. MISSING DATA ANALYSIS AND MINIMIZING BIAS

2b7.1. Describe the method of testing conducted to identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias (*describe the steps—do not just name a method; what statistical analysis was used*)

The data source for this measure is claims data. Due to the nature of claims data (i.e., for billing purposes), there is typically very little missing data; further, it is difficult to ascertain when claims data is or is not missing.

2b7.2. What is the overall frequency of missing data, the distribution of missing data across providers, and the results from testing related to missing data? (*e.g., results of sensitivity analysis of the effect of various rules for missing data/nonresponse; if no empirical sensitivity analysis, identify the approaches for handling missing data that were considered and pros and cons of each*)

Not addressed due to the nature of claims data.

2b7.3. What is your interpretation of the results in terms of demonstrating that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias**?** (i.*e., what do the results mean in terms of supporting the selected approach for missing data and what are the norms for the test conducted; if no empirical analysis, provide rationale for the selected approach for missing data*)

Not applicable.

References

Adams JL, Mehrotra A, Thomas JW, McGlynn EA (2010). Physician cost profiling-reliability and risk of misclassification. New England Journal of Medicine; 362:1014–1021.

Adams JL (2009). The Reliability of Provider Profiling: A Tutorial. Rand Health, Santa Monica, CA. Last accessed online at: http://www.rand.org/pubs/technical_reports/TR653.

Eijkenaar F, van Vliet RC. (2013). Profiling individual physicians using administrative data from a single insurer: variance components, reliability, and implications for performance improvement efforts. Med Care. 51(8):731-9

Ene, M, Leighton, EA, Blue, GL, Bell, BA (2015). Models for categorical data using SAS PROC GLIMMIX. Sas Global Forum, Dallas TX, April 26-29, Paper 3430. Last accessed online at: http://support.sas.com/resources/papers/proceedings15/3430-2015.pdf.

Fung, V, Schmittdiel, JA, Fireman, B, Meer, A, Thomas, S, Hsu, J, Selby, JV (2010). Meaningful variation in performance: A systematic literature review. Medical Care, 48(2): 140-148.

He, Y, Selck, F, and Normand, SL (2014). On the accuracy of classifying hospitals on their performance measures. Statistics in Medicine; 33: 1081-1103.

Martin JA, Hamilton BE, Osterman MJ, Curtin SC, Matthews TJ. (2015) Births: final data for 2013. Natl Vital Stat Rep; 64(1):1-65.

Molenberghs G, Laenen A, Vengeneugden T (2007). Estimating reliability and generalizability from hierarchical biomedical data. Journal of Biopharmaceutical Statistics, 17(4): 595-627.

National Committee for Quality Assurance (NCQA). HEDIS 2015: Healthcare Effectiveness Data and Information Set. Vol. 1, narrative. Washington (DC): National Committee for Quality Assurance (NCQA); Retrieved from: http://www.qualitymeasures.ahrq.gov/content.aspx?id=48621

NQF 2009. National Quality Forum, Health Information Technology Expert Panel II- Health IT Enablement of Quality Measurement. Washington DC.

Raudenbush SW and Bryk AS. *Hierarchical Linear Models: Applications and Data Analysis Methods (2nd Edition)*. Sage Publications, 2002: Thousand Oaks.

Ridout MS, Demetrio CGB, Firth D (1999). Estimating intraclass correlation for binary data. Biometrics, 55(1): 137-148.

Romero L, Pazol K, Warner L, Gavin L, Moskosky S, Besera G, et al. (2015). Vital Signs: Trends in Use of Long-Acting Reversible Contraception Among Teens Aged 15–19 Years Seeking Contraceptive Services — United States, 2005–2013. MMWR Weekly Report April 10, 2015 / 64(13);363-369

Scholle SH, Roski J, Adams JL, Dunn DL, Kerr EA, Pilleterre-Dugan D, Jensen RE (2008). Benchmarking physician performance: reliability of individual and composite measures. Am J of Managed Care, 14(12): 833-838.

Sedgh G, Finer LB, Bankole A, Eilers MA, Singh S. (2015) Adolescent pregnancy, birth, and abortion rates across countries: levels and recent trends. J Adolesc Health; 56(2):223-30.

Woltman H, Feldstain A, MacKay JC, Rocchi M. (2012). An introduction to hierarchical linear modeling. Tutorials in Quantitative Methods for Psychology; 8(1):52–69.

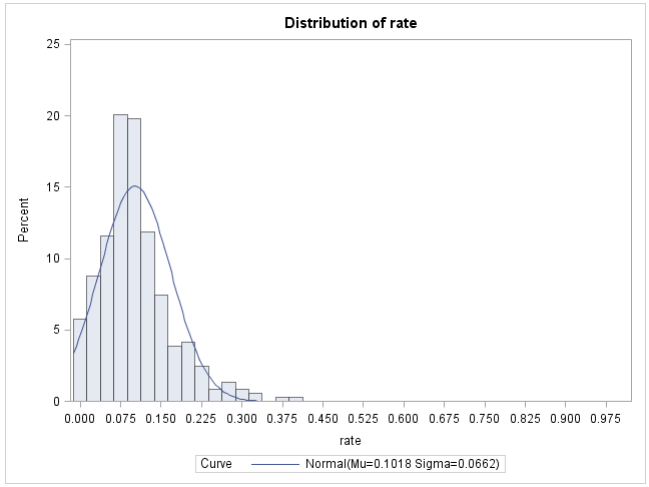
Wu S, Crespi CM, Wong WK (2012). Comparison of methods for estimating the intraclass correlation coefficient for binary responses in cancer prevention cluster randomized trials. Contemp Clin Trials, 33(5): 869-880.

Table 1. Rates and reliabilities for use of LARC methods, 25 PPFA affiliates, 2014

Affiliate ID	LARCmeasure: 15 to <21 Years				LARCmeasure: 21 to 45 years				LARCmeasure: all age groups			
	Used LARC	Total N	Rate	HC Within Affiliate Reliability	Used LARC	Total N	Rate	HC Within Affiliate Reliability	Used LARC	Total N	Rate	HC Within Affiliate Reliability
1	1516	7869	0.193	**0.9896**	5464	29638	0.184	**0.9976**	6980	37507	0.186	**0.9982**
3	1878	26591	0.071	**0.9855**	9687	88881	0.109	**0.9959**	11565	115472	0.100	**0.9970**
4	638	4147	0.154	**0.9832**	3073	21430	0.143	**0.9994**	3711	25577	0.145	**0.9994**
5	4979	42698	0.117	**0.9827**	18747	131187	0.143	**0.9937**	23726	173885	0.136	**0.9959**
6	273	2651	0.103	**0.9243**	907	7362	0.123	**0.9741**	1180	10013	0.118	**0.9825**
9	2035	25268	0.081	**0.9949**	10757	88455	0.122	**0.9978**	12792	113723	0.112	**0.9984**
10	1753	15188	0.115	**0.9408**	5839	47698	0.122	**0.9606**	7592	62886	0.121	**0.9773**
12	552	4839	0.114	**0.9557**	1181	10209	0.116	**0.9818**	1733	15048	0.115	**0.9861**
37	161	1965	0.082	**0.8686**	450	4194	0.107	**0.8263**	611	6159	0.099	**0.9194**
38	452	6093	0.074	**0.9484**	1307	10645	0.123	**0.9387**	1759	16738	0.105	**0.9685**
40	566	5030	0.113	**0.9356**	1336	10843	0.123	**0.9487**	1902	15873	0.120	**0.9690**
41	575	5466	0.105	**0.8980**	2448	17562	0.139	**0.9576**	3023	23028	0.131	**0.9639**
44	1102	11489	0.096	**0.9904**	3591	33620	0.107	**0.9952**	4693	45109	0.104	**0.9967**
47	349	5644	0.062	**0.9897**	876	16648	0.053	**0.9848**	1225	22292	0.055	**0.9931**
53	1032	8741	0.118	**0.9738**	2984	28791	0.104	**0.9852**	4016	37532	0.107	**0.9908**
54	656	3122	0.210	**0.8474**	1313	6614	0.199	**0.8879**	1969	9736	0.202	**0.9304**
59	371	3682	0.101	**0.6688**	906	9778	0.093	**0.8604**	1277	13460	0.095	**0.8329**
60	61	436	0.140	**0.0000**	163	1265	0.129	**0.0000**	224	1701	0.132	**0.0000**
70	145	4154	0.035	**0.9869**	454	12436	0.037	**0.9990**	599	16590	0.036	**0.9993**
73	57	996	0.057	**0.9469**	191	2825	0.068	**0.9704**	248	3821	0.065	**0.9784**
75	98	1171	0.084	**0.9895**	442	5070	0.087	**0.9930**	540	6241	0.087	**0.9950**
76	109	3817	0.029	**0.9762**	767	11648	0.066	**0.9803**	876	15465	0.057	**0.9876**
77	781	11359	0.069	**0.9553**	2329	31393	0.074	**0.9799**	3110	42752	0.073	**0.9885**
79	84	1260	0.067	**0.9760**	143	5149	0.028	**0.7775**	227	6409	0.035	**0.9796**
81	24	294	0.082	**0.0000**	134	1561	0.086	**0.0000**	158	1855	0.085	**0.0000**
Total or Mean	**20247**	**203970**	**0.099**		**75489**	**634902**	**0.119**		**95736**	**838872**	**0.114**	
		σ Level 2	**ICC**	**Overall Affiliate Reliability**		**σ Level 2**	**ICC**	**Overall Affiliate Reliability**		**σ Level 2**	**ICC**	**Overall Affiliate Reliability**
Reliability using Median Affiliate Patient Volume	**Median n** **4839**	**0.2374**	**0.0673**	**0.9971**	**Median n** **11648**	**0.2381**	**0.0675**	**0.9988**	**Median n** **16590**	**0.2163**	**0.0617**	**0.9991**
Reliability using Minimum Patient Volume (Floor)	Min n 294	0.2374	0.0673	0.9550	Min n 1265	0.2381	0.0675	0.9892	Min n 1701	0.2163	0.0617	0.9911

Table 2. Distributions of rates and ICCs among health centers (n=363) for use of LARC methods among 25 PPF affiliates. 2014

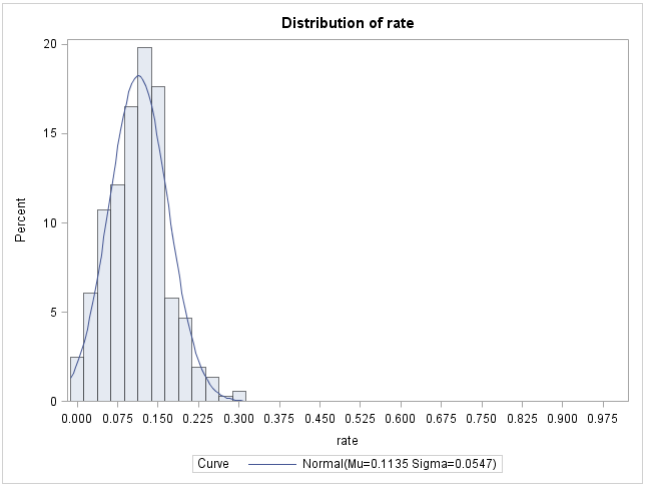
LARCmeasure: 15 to <21 Years	LARCmeasure: 21 to 45 years	



	cases (n)	rate
Mean	561.9	0.10
Median	366	0.09
SD	552.3	0.07
Variance	305043	0.00
Range	2976	0.39
Interquartile	629	0.07
Median ICC		0.06

HCs with rate <.02 = 32

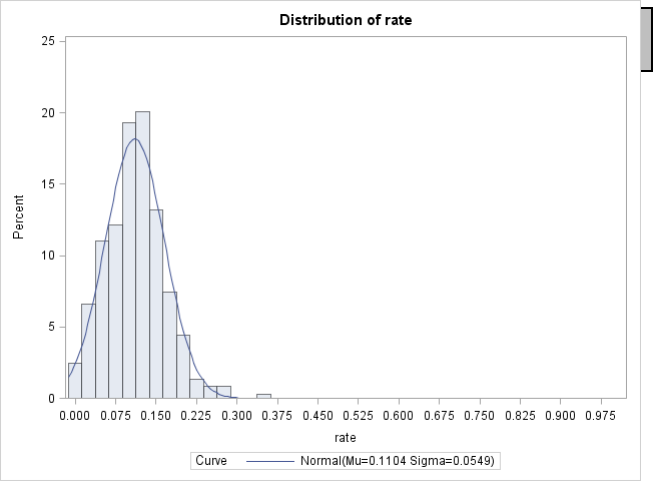
Quantile	cases (n)	rate
100% Max	2984	0.39
95%	1766	0.23
90%	1410	0.19
75% Q3	787	0.13
50% Med	366	0.09
25% Q1	158	0.06
10%	83	0.02
5%	53	0.01
0% Min	8	0.00



	cases (n)	rate
Mean	1749.0	0.11
Median	1016	0.12
SD	1909	0.05
Variance	3645550	0.00
Range	11360	0.31
Interquartile	2145	0.07
Median ICC		0.04

HCs with rate <.02 = 11

Quantile	cases (n)	rate
100% Max	11391	0.31
95%	5489	0.21
90%	4544	0.18
75% Q3	2516	0.15
50% Med	1016	0.12
25% Q1	371	0.08
10%	149	0.04
5%	92	0.03
0% Min	31	0.00



	cases (n)	rate
Mean	2310.9	0.11
Median	1379	0.11
SD	2424	0.05
Variance	5875321	0.00
Range	13287	0.35
Interquartile	2757	0.07
Median ICC		0.05

HCs with rate <.02 = 24

Quantile	cases (n)	rate
100% Max	13335	0.35
95%	7198	0.21
90%	5872	0.18
75% Q3	3315	0.14
50% Med	1379	0.11
25% Q1	558	0.07
10%	240	0.04
5%	141	0.03
0% Min	48	0.00

Table 3. Rates and reliabilities for use of LARC method, Iowa Medicaid Enterprise, 2013, by region and type of benefit

Public Health Region	LARCMeasure: 15 to <21 Years						LARCMeasure: 21 to 45 years						LARCMeasure: all age groups					
	Not Used	Used LARC	Total N	Rate			Not Used	Used LARC	Total N	Rate			Not Used	Used LARC	Total N	Rate		
1	3204	256	3460	0.074			8715	873	9588	0.091			11919	1129	13048	0.087		
2	1034	120	1154	0.104			2577	329	2906	0.113			3611	449	4060	0.111		
3	1096	80	1176	0.068			2851	324	3175	0.102			3947	404	4351	0.093		
4	992	95	1087	0.087			2635	252	2887	0.087			3627	347	3974	0.087		
5	1566	135	1701	0.079			3966	393	4359	0.090			5532	528	6060	0.087		
6	2815	306	3121	0.098			9132	1004	10136	0.099			11947	1310	13257	0.099		
Total or Mean	**10707**	**992**	**11699**	**0.085**			**29876**	**3175**	**33051**	**0.096**			**40583**	**4167**	**44750**	**0.093**		
					Var L1	**ICC**	Region Reliability (Var L1)				**Var L1**	**ICC**	Region Reliability (Var L1)				**Var L1**	**ICC**	Region Reliability (Var L1)
Median Patient Volume Among Affiliates		Median	1438.5	0.018227	0.0055		**0.8887**		Median	3767	0.000567	0.0017	**0.8666**		Median	5205.5	0.00724	0.0022	**0.9197**
Minimum Patient Volume (Floor)		Min	1087	0.018227	0.0055		0.8579		Min	2887	0.000567	0.0017	0.8327		Min	3974	0.00724	0.0022	0.8974

Type of Benefit	LARCMeasure: 15 to <21 Years						LARCMeasure: 21 to 45 years						LARCMeasure: all age groups					
	Not Used	Used LARC	Total N	Rate			Not Used	Used LARC	Total N	Rate			Not Used	Used LARC	Total N	Rate		
Family Planning Waiver	5698	747	6445	0.116			20880	2688	23568	0.114			26578	3435	30013	0.114		
Non-Family Planning Waiver	5009	245	5254	0.047			8996	487	9483	0.051			14005	732	14737	0.050		
Total or Mean	**10707**	**992**	**11699**	**0.085**			**29876**	**3175**	**33051**	**0.096**			**40583**	**4167**	**44750**	**0.093**		
					Var L2	**ICC**	Benefit type Reliability (Var L2)				**Var L2**	**ICC**	Benefit type Reliability (Var L2)				**Var L2**	**ICC**	Benefit type Reliability (Var L2)
Reliability Based on Median Patient Volume Among Health Centers		Median	5849.5	0.2408	0.0682		**0.9977**		Median	16525.5	0.1867	0.0537	**0.9989**		Median	22375	0.2043	0.0585	**0.9993**
Calculated Based on Minimum Patient Volume (Floor)		Min	5254	0.2408	0.0682		0.9974		Min	9483	0.1867	0.0537	0.9981		Min	14737	0.2043	0.0585	0.9989

Region 1/Family Planning Waiver	1464	72	1536	0.047			2421	123	2544	0.048			3885	195	4080	0.048		
Region 1/Non-Family Planning Waiver	1740	184	1924	0.096			6294	750	7044	0.106			8034	934	8968	0.104		
Region 2/Family Planning Waiver	438	25	463	0.054			617	38	655	0.058			1055	63	1118	0.056		
Region 2/Non-Family Planning Waiver	596	95	691	0.137			1960	291	2251	0.129			2556	386	2942	0.131		
Region 3/Family Planning Waiver	588	21	609	0.034			726	49	775	0.063			1314	70	1384	0.051		
Region 3/Non-Family Planning Waiver	508	59	567	0.104			2125	275	2400	0.115			2633	334	2967	0.113		
Region 4/Family Planning Waiver	511	25	536	0.047			857	44	901	0.049			1368	69	1437	0.048		
Region 4/Non-Family Planning Waiver	481	70	551	0.127			1778	208	1986	0.105			2259	278	2537	0.110		
Region 5/Family Planning Waiver	702	39	741	0.053			1423	74	1497	0.049			2125	113	2238	0.050		
Region 5/Non-Family Planning Waiver	864	96	960	0.100			2543	319	2862	0.111			3407	415	3822	0.109		
Region 6/Family Planning Waiver	1306	63	1369	0.046			2952	159	3111	0.051			4258	222	4480	0.050		
Region 6/Non-Family Planning Waiver	1509	243	1752	0.139			6180	845	7025	0.120			7689	1088	8777	0.124		
Total or Mean	**10707**	**992**	**11699**	**0.085**			**29876**	**3175**	**33051**	**0.096**			**40583**	**4167**	**44750**	**0.093**		
				Var L2	**ICC**	**Region by benefit type Reliability (Var L2)**				**Var L2**	**ICC**	**Region by benefit type Reliability (Var L2)**				**Var L2**	**ICC**	**Region by benefit type Reliability (Var L2)**
Reliability Based on Median Patient Volume Among Health Centers		Median	716	0.2537	0.0716	**0.98922**		Median	2325.5	0.1775	0.0512	**0.99021**		Median	2954.5	0.2003	0.0574	**0.99455**
Calculated Based on Minimum Patient Volume (Floor)		Min	463	0.2537	0.0716	0.97728		Min	655	0.1775	0.0512	0.97225		Min	1118	0.2003	0.0574	0.98555

Table 4. Rates and reliabilities for use of LARC method, Wisconsin Medicaid, 2014, by health plan/HMO

HMO	LARCMeasure: 15 to <21 Years						LARCMeasure: 21 to 45 years						LARCMeasure: all age groups					
	Not Used	Used LARC	Total N	Rate			Not Used	Used LARC	Total N	Rate			Not Used	Used LARC	Total N	Rate		
1	4598	234	4832	0.048			12894	1149	14043	0.082			17492	1383	18875	0.073		
2	1742	96	1838	0.052			5314	374	5688	0.066			7056	470	7526	0.062		
3	861	59	920	0.064			2633	229	2862	0.080			3494	288	3782	0.076		
4	1682	113	1795	0.063			5188	493	5681	0.087			6870	606	7476	0.081		
5	1147	84	1231	0.068			3673	263	3936	0.067			4820	347	5167	0.067		
6	203	16	219	0.073			662	63	725	0.087			865	79	944	0.084		
7	518	40	558	0.072			1475	133	1608	0.083			1993	173	2166	0.080		
8	326	26	352	0.074			1001	95	1096	0.087			1327	121	1448	0.084		
9	1539	84	1623	0.052			5767	397	6164	0.064			7306	481	7787	0.062		
10	572	46	618	0.074			1524	159	1683	0.094			2096	205	2301	0.089		
11	4621	277	4898	0.057			13996	1170	15166	0.077			18617	1447	20064	0.072		
12	1167	72	1239	0.058			4027	263	4290	0.061			5194	335	5529	0.061		
13	246	23	269	0.086			775	78	853	0.091			1021	101	1122	0.090		
14	2009	140	2149	0.065			5168	428	5596	0.076			7177	568	7745	0.073		
15	52	4	56	0.071			226	14	240	0.058			278	18	296	0.061		
16	4860	254	5114	0.050			17460	1415	18875	0.075			22320	1669	23989	0.070		
17	517	42	559	0.075			1346	187	1533	0.122			1863	229	2092	0.109		
Total or Mean	**26660**	**1610**	**28270**	**0.057**			**83129**	**6910**	**90039**	**0.077**			**109789**	**8520**	**118309**	**0.072**		
				Var L1	**ICC**	Overall HMO Reliability (Var L1)				**Var L1**	**ICC**	Overall HMO Reliability (Var L1)				**Var L1**	**ICC**	Overall HMO Reliability (Var L1)
Median Patient Volume Among Affiliates		Median n	1231	0.01418	0.0043	**0.8414**		Median n	3936	0.02718	0.0082	**0.9702**		Median n	5167	0.02218	0.0067	**0.9721**
Minimum Patient Volume (Floor)		Min n	56	0.01418	0.0043	0.1944		Min n	240	0.02718	0.0082	0.6647		Min n	296	0.02218	0.0067	0.6662

3. Feasibility

Extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

3a. Byproduct of Care Processes

For clinical measures, the required data elements are routinely generated and used during care delivery (e.g., blood pressure, lab test, diagnosis, medication order).

3a.1. **Data Elements Generated as Byproduct of Care Processes.**

Coded by someone other than person obtaining original information (e.g., DRG, ICD-9 codes on claims)

If other:

3b. Electronic Sources

The required data elements are available in electronic health records or other electronic sources. If the required data are not in electronic health records or existing electronic sources, a credible, near-term path to electronic collection is specified.

3b.1. **To what extent are the specified data elements available electronically in defined fields?** *(i.e., data elements that are needed to compute the performance measure score are in defined, computer-readable fields)*

ALL data elements are in defined fields in electronic claims

3b.2. **If ALL the data elements needed to compute the performance measure score are not from electronic sources, specify a credible, near-term path to electronic capture, OR provide a rationale for using other than electronic sources.**

3b.3. **If this is an eMeasure, provide a summary of the feasibility assessment in an attached file or make available at a measure-specific URL.**

Attachment:

3c. Data Collection Strategy

Demonstration that the data collection strategy (e.g., source, timing, frequency, sampling, patient confidentiality, costs associated with fees/licensing of proprietary measures) can be implemented (e.g., already in operational use, or testing demonstrates that it is ready to put into operational use). For eMeasures, a feasibility assessment addresses the data elements and measure logic and demonstrates the eMeasure can be implemented or feasibility concerns can be adequately addressed.

3c.1. **Describe what you have learned/modified as a result of testing and/or operational use of the measure regarding data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, time and cost of data collection, other feasibility/implementation issues.**

IF a PRO-PM, consider implications for both individuals providing PROM data (patients, service recipients, respondents) and those whose performance is being measured.

The measure has been piloted in Iowa and Wisconsin Medicaid programs, PPFA, and is currently being used by 13 state Medicaid programs funded by the Center for Medicaid and CHIP Services (CMCS) to report on the measure as part of the CMCS Maternal and Infant Health initiative (see below). Overall, these experiences have confirmed that the measures can be feasibly and meaningfully calculated using existing claims data. Some other lessons learned are summarized below:

- Feasibility and biases introduced by using a 'look back' period. Several states have inquired about the use of a 'look back' period to identify women who were sterilized or received LARC in a year preceding the measurement year, or who had a procedure that caused them to become infecund for non-contraceptive reasons (e.g., oophorectomy). We are open to exporing the feasibility of this, but several challenges related to feasibility and potential biases caused us to decide against including a look-back period in this version of the measure: (1) LARC last for a period 3-10 years, so an extensive period of look-back would be needed for this purpose as well; and (2) high rates of moving on and off a health plan may make it very hard to interpret the results of a look-back, especially if only a small percentage of currently enrolled women were enrolled for the entire look-back period.
- State-specific NDC codes may need to be added: Several of the CMCS-funded states found that some NDC codes were missing from the measure specifications, especially for oral contraception. Obtaining a comprehensive list of NDC codes for contraception is challenging, and we relied on the list compiled by NCQA for the calculation of the HEDIS chlamydia screening measure. We have asked states to include the additional NDC codes when they calculate the measure, but to report to us the additional codes at the time of submission. This is consistent with the approach used by NCQA for the HEDIS chlamydia screening measure.
- With this measure, the imprecisions described above are not a major issue because we only trying to identify very low

performers and not necesarily set a high benchmark.
Existing administrative claims data has several limitation with regard to the measurement of unintended pregnancy. In particular, the data does not capture the client's history of sexual experience, their desire to become pregnant, or LARC insertion in a year preceding the measurement year. In the medium-term, these limitations are not a serious concern because there is so much room for improvement. Yet for the longer term, OPA is actively pursuing efforts to develop either a hybrid or all-electronic version of the measure. We are piloting a hybrid measure in two community health centers in the coming year, and are working with Integrating the Healthcare Enterprise to create a family planning profile that includes all data elements needed to compute the measures with electronic data. If these plans are successful, we expect to be able to submit an application for either a hybrid or e-measure version of the measure within a 3-year period of time.

3c.2. **Describe any fees, licensing, or other requirements to use any aspect of the measure as specified** *(e.g., value/code set, risk model, programming code, algorithm)*.
Not applicable.

4. Usability and Use

Extent to which potential audiences (e.g., consumers, purchasers, providers, policy makers) are using or could use performance results for both accountability and performance improvement to achieve the goal of high-quality, efficient healthcare for individuals or populations.

4a. Accountability and Transparency

Performance results are used in at least one accountability application within three years after initial endorsement and are publicly reported within six years after initial endorsement (or the data on performance results are available). If not in use at the time of initial endorsement, then a credible plan for implementation within the specified timeframes is provided.

4.1. Current and Planned Use

NQF-endorsed measures are expected to be used in at least one accountability application within 3 years and publicly reported within 6 years of initial endorsement in addition to performance improvement.

Planned	Current Use (for current use provide URL)
Public Reporting	Quality Improvement (Internal to the specific organization) CMCS Maternal and Infant Health Initiative https://www.medicaid.gov/Medicaid-CHIP-Program-Information/By-Topics/Quality-of-Care/Maternal-and-Infant-Health-Care-Quality.html. OPA Title X family planning program www.familyplanningdashboard.com

4a.1. For each CURRENT use, checked above, provide:

- Name of program and sponsor
- Purpose
- Geographic area and number and percentage of accountable entities and patients included

The Center for Medicaid and CHIP Services

As part of the CMCS Maternal and Infant Health Initiative, state Medicaid programs will report on the measure on a voluntary basis. Nearly two out of every three adult women enrolled in Medicaid are in their reproductive years (ages 19-44), and Medicaid currently finances about 45% of all births in the United States. Recognizing this, the Centers for Medicaid and CHIP Services (CMCS) has developed the Maternal and Infant Health initiative to improve the quality of maternity care, birth outcomes and in measuring how care is delivered to women. CMCS will collect and report data on the new contraceptive use measure on a developmental basis, to help states track the use of most and moderately effective methods of contraception and to drive changes in care practices and delivery. The measure specifications developed for use in the Maternal and Infant Health initiative are found in Appendix C. For more information about the Maternal and Infant Health initiative, see: https://www.medicaid.gov/Medicaid-CHIP-Program-Information/By-Topics/Quality-of-Care/Maternal-and-Infant-Health-Care-Quality.html.

OPA's Title X Program

In 2014, the Title X program funded 94 grantees that support a network of 4100 service sites, which in turn served 4.1 million clients. Approximately 75% of US counties have at least one Title X-funded clinic. In addition, in 2015-2016, OPA has developed an online contraceptive measures calculator for use throughout its Title X program. The calculator enables grantees, using FPAR data, to automatically calculate the two measures for each site in its network. OPA is also piloting a learning collaborative, based on the Institute of Healthcare Improvement's Breakthrough Series model, to support grantees to assess and optimize their performance on the access and outcome measures through employing strategies documented in an evidence-based change package. The learning collaborative involves coaching and supporting the members through the plan, do, study, act cycle for selected change package strategies. The collaborative includes an on-line community of practice to further promote peer exchange and learning. The collaborative serves as a model that can be spread throughout the Title X program and adapted in other settings. The change package and other materials from the collaborative, once tested, will be made available for use by the full reproductive health community. As an example, the Iowa Department of Public Health (IDPH) is a Title X grantee that has pioneered use of this performance measure within its service site network, and provided the following comments on the usability of the measure: "IDPH has used the data to assess access to most and moderately effective contraceptive methods in the Title X provider network; explored barriers/challenges and opportunities to improve access; identified training needs for grantees and strategies to improve performance. The data has provided IDPH a standardized way of looking at access to contraceptive methods, how services are being provided and how to identify gaps in services. The measure will give us an opportunity to consider why we are seeing the results we are seeing, to think about those results and determine if and where changes can be made to improve outcomes. We have used the

data to help develop the competitive application, and will use it in years 2-5 of the Title X award period as a variable in the funding formula."

Planned Parenthood Federation of America
In 2014, Planned Parenthood of America (PPFA) comprised 66 independently incorporated affiliates, operating approximately 700 health centers in the United States, providing reproductive health care to nearly 2.7 million patients. In August 2012, Planned Parenthood Federation of America (PPFA) added a Clinical Quality Improvement (CQI) Department to its Health Care Division. The CQI Department coordinates a federation-wide clinical quality improvement program working with Affiliates. As part of the federation-wide quality improvement program and to capitalize on this investment in health IT, a set of core reports were built as key measures of quality of care and health outcomes including use of contraception services. Nearly 70% of the affiliates partner with the PPFA CQI Department and receive quality reports on key clinical measures as well as technical assistance for quality improvement activities.

4a.2. **If not currently publicly reported OR used in at least one other accountability application (e.g., payment program, certification, licensing) what are the reasons?** (*e.g., Do policies or actions of the developer/steward or accountable entities restrict access to performance results or impede implementation?*)

4a.3. **If not currently publicly reported OR used in at least one other accountability application, provide a credible plan for implementation within the expected timeframes -- any accountability application within 3 years and publicly reported within 6 years of initial endorsement.** (*Credible plan includes the specific program, purpose, intended audience, and timeline for implementing the measure within the specified timeframes. A plan for accountability applications addresses mechanisms for data aggregation and reporting.*)
1) The Measure Applications Partnership (MAP) recommended inclusion of these measures in the CMCS core adult and child measure set, pending NQF endorsement.
2) OPA's Title X program intends to make publicly available the performance of all grantees (individually and aggregated by state) within a 3 year period of time.

4b. Improvement

Progress toward achieving the goal of high-quality, efficient healthcare for individuals or populations is demonstrated. If not in use for performance improvement at the time of initial endorsement, then a credible rationale describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.

4b.1. **Progress on Improvement. (Not required for initial endorsement unless available.)**
Performance results on this measure (current and over time) should be provided in 1b.2 and 1b.4. Discuss:

- **Progress (trends in performance results, number and percentage of people receiving high-quality healthcare)**
- **Geographic area and number and percentage of accountable entities and patients included**

4b.2. **If no improvement was demonstrated, what are the reasons? If not in use for performance improvement at the time of initial endorsement, provide a credible rationale that describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.**

4c. Unintended Consequences

The benefits of the performance measure in facilitating progress toward achieving high-quality, efficient healthcare for individuals or populations outweigh evidence of unintended negative consequences to individuals or populations (if such evidence exists).

4c.1. **Were any unintended negative consequences to individuals or populations identified during testing; OR has evidence of unintended negative consequences to individuals or populations been reported since implementation? If so, identify the negative unintended consequences and describe how benefits outweigh them or actions taken to mitigate them.**
No unintended negative consequences were identified. The one issue that has arisen as a potential concern is that the measure may lead to coercive practices in which women are not offered a free choice of methods and are pressured into using a LARC method (Gold 2014, Dehlendorf 2015). We do not think this will be a concern for two reasons:

- Although existing research (Harper 2015, Winner 2012) show a high percentage of women will choose LARC when given the opportunity, the focus of this measure is on ensuring access to these methods by monitoring very low rates of use (e.g., below 2%, or

well below the median of all reporting units). Further, we have explicitly proposed that this measure should not have a benchmark encouraging high rates of use, and that it would be an inappropriate measure to use in pay-for-performance or similar programs. If the measure is used as intended (i.e., to assess lack of access), this should remove pressure on providers to inappropriately "promote" LARC methods.

- CDC-OPA recommendations describe in detail how to provide client-centered, non-coercive contraceptive counseling, and efforts to support use of the measure should be accompanied by efforts to increase awareness of the CDC-OPA recommendations (CDC/OPA 2014). Further, OPA has funded the development of training on how to provide client-centered counseling, which is available to all providers on the OPA-supported training website (www.fpntc.org).

References:

- Dehlendorf C, Bellanca H, Policar M (2015). Performance measures for contraceptive care: what are we actually trying to measure? Contraception. Jun;91(6):433-7.
- Gold R (2014). Guarding Against Coercion While Ensuring Access: A Delicate Balance. Guttmacher Policy Review. Summer 2014, Volume 17, Number 3.
- Harper C, Rocca CH, Thompson KM, Morfesis J, Goodman S, Darney PD, Westhoff CL, Speidel JJ (2015). Reductions in pregnancy rates in the USA with long-acting reversible contraception: a cluster randomised trial. Lancet. Volume 386, No. 9993, p562–568, 8 August 2015.
- Winner B, Peipert J, Qiuhong Z, Buckel C, Madden T et al (2012). Effectiveness of Long-Acting Reversible Contraception, The New England Journal of Medicine, 366 (21): 1998-2007.
- CDC/OPA (2014). Providing Quality Family Planning Services: Recommendations of CDC and the US Office of Population Affairs, MMWR Recommendations and Reports, April 24, 2014.
- OPA-funded training materials are available at this website: www.fpntc.org

5. Comparison to Related or Competing Measures

If a measure meets the above criteria and there are endorsed or new related measures (either the same measure focus or the same target population) or competing measures (both the same measure focus and the same target population), the measures are compared to address harmonization and/or selection of the best measure.

5. Relation to Other NQF-endorsed Measures
Are there related measures (conceptually, either same measure focus or target population) or competing measures (conceptually both the same measure focus and same target population)? If yes, list the NQF # and title of all related and/or competing measures.
No

5.1a. **List of related or competing measures (selected from NQF-endorsed measures)**

5.1b. **If related or competing measures are not NQF endorsed please indicate measure title and steward.**

5a. Harmonization

The measure specifications are harmonized with related measures;
OR
The differences in specifications are justified

5a.1. **If this measure conceptually addresses EITHER the same measure focus OR the same target population as NQF-endorsed measure(s):**
Are the measure specifications completely harmonized?

5a.2. **If the measure specifications are not completely harmonized, identify the differences, rationale, and impact on interpretability and data collection burden.**

5b. Competing Measures

The measure is superior to competing measures (e.g., is a more valid or efficient way to measure);
OR
Multiple measures are justified.

5b.1. **If this measure conceptually addresses both the same measure focus and the same target population as NQF-endorsed measure(s):**
Describe why this measure is superior to competing measures (e.g., a more valid or efficient way to measure quality); OR provide a rationale for the additive value of endorsing an additional measure. (Provide analyses when possible.)
NOTE: OPA is submitting two other applications for NQF endorsement, which are complementary to this measure application. One of the applications focuses on use of most and moderately effective contraceptive methods in a key sub-population of women at risk of unintended pregnancy, i.e., postpartum women. The other application focuses on use of most (sterilization, IUD, implant) and moderately (shot, pill, patch, ring, diaphragm) effective methods of contraception, of which LARC methods are a subset.

Appendix

A.1 Supplemental materials may be provided in an appendix. All supplemental materials (such as data collection instrument or methodology reports) should be organized in one file with a table of contents or bookmarks. If material pertains to a specific submission form number, that should be indicated. Requested information should be provided in the submission form and required attachments. There is no guarantee that supplemental materials will be reviewed.

Attachment **Attachment:** nqf_Appendix_LARC.pdf

Contact Information

Co.1 **Measure Steward (Intellectual Property Owner):** US Office of Population Affairs
Co.2 **Point of Contact:** Loretta, Gavin, loretta.gavin@hhs.gov, 240-453-2826-
Co.3 **Measure Developer if different from Measure Steward:** US Office of Population Affairs
Co.4 **Point of Contact:** Loretta, Gavin, loretta.gavin@hhs.gov, 240-453-2826-

Additional Information

Ad.1 **Workgroup/Expert Panel involved in measure development**
Provide a list of sponsoring organizations and workgroup/panel members' names and organizations. Describe the members' role in measure development.
The formal steward of the measure is OPA, but the measure is also supported by the U.S. Centers for Disease Control and Prevention, Division of Reproductive Health (http://www.cdc.gov/reproductivehealth/index.html).
Representatives of numerous organizations were involved in helping to develop the measure. Their roles included helping define the conceptual basis and rationale for the measure, piloting and testing data, reviewing draft measure specifications, and/or reviewing the NQF application.

- U.S. Office of Population Affairs: Susan Moskosky MS
- U.S. Centers for Disease Control and Prevention: Lisa Romero PhD, Cheryl Robbins PhD, Peter Briss MD, Karen Pazol PhD, Maria Rivera MPH
- U.S. Center for Medicaid and CHIP Services: Lekisha Daniel-Robinson
- Planned Parenthood Federation of America: Jennifer Fuld PhD, Carolyn Westhoff MD, Mytri Singh MPH, Joycelyn Benson MPH, Kellan Smith
- Far Harbor LLC: Philip A. Hastings, PhD and Prasant Mohanty, MBBS, MPH
- Iowa Department of Public Health and Iowa Medicaid Enterprise: Denise Wheeler MS, Debra Kane PhD, Brittni Frederiksen PHD, Mikki Stier, Julie Lovelady, Dr. Jason Kessler, Sally Nadolsky (retired), Mark McMahon, Gerd Clabaugh, Brenda Dobson, Marcus Johnson-Miller, Abigail Holicky, Jessica Riggs
- Wisconsin Department of Health Services: Crystal Gibson MPH, Angela Rohan PhD
- Illinois Medicaid: Julie Doetsch, Gwen Smith
- National Contraceptive Quality Measures Workgroup

Measure Developer/Steward Updates and Ongoing Maintenance
Ad.2 **Year the measure was first released:**
Ad.3 **Month and Year of most recent revision:**
Ad.4 **What is your frequency for review/update of this measure?**
Ad.5 **When is the next scheduled review/update for this measure?**

Ad.6 **Copyright statement:** Not applicable.
Ad.7 **Disclaimers:** Not applicable.

Ad.8 **Additional Information/Comments:** Not applicable.