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

Purple text represents the responses from measure developers.

Red text denotes developer information that has changed since the last measure evaluation review.

Brief Measure Information

NQF #: 0470

Corresponding Measures:

De.2. Measure Title: Incidence of Episiotomy

Co.1.1. Measure Steward: Christiana Care Health System

De.3. Brief Description of Measure: Percentage of vaginal deliveries (excluding those coded with shoulder dystocia) during which an episiotomy is performed.

1b.1. 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 fasciitis. Predicated on these concerns, ACOG has called for "restricted use of episiotomy".

S.4. 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.

S.6. Denominator Statement: All vaginal deliveries during the analytic period- monthly, quarterly, yearly etc. excluding those coded with a shoulder dystocia ICD-10: O66.0).

S.8. 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.

De.1. Measure Type: Process

S.17. Data Source: Claims, Electronic Health Data, Electronic Health Records, Paper Medical Records

S.20. Level of Analysis: Facility

IF Endorsement Maintenance – Original Endorsement Date: Oct 24, 2008 **Most Recent Endorsement Date:** Oct 25, 2016

IF this measure is included in a composite, NQF Composite#/title:

IF this measure is paired/grouped, NQF#/title:

De.4. IF PAIRED/GROUPED, what is the reason this measure must be reported with other measures to appropriately interpret results? NA

Preliminary Analysis: Maintenance of Endorsement

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 *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. For measures derived from patient report, evidence also should demonstrate that the target population values the measured process or structure and finds it meaningful.

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 of prior review in 2016

- This process measure was last reviewed in 2016. The developer reported that this measure is intended to reduce the incidence of episiotomy during vaginal delivery, thereby reducing rates of perineal injury.
- The 2016 evidence focused on a recommendation from an April 2006 American College of Obstetricians and Gynecologists (ACOG) Practice Bulletin (no. 71). This evidence received an A grade from ACOG.

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 cited a new ACOG practice bulletin (no. 165) from July 2016. This updated evidence provides further evidence that routine use of episiotomy may be detrimental to the mother and is not beneficial.
- The evidence cited by the developer does not describe an optimal episiotomy level. However, the developer reports data from 2014 from within their facilities: "6-7% of women continue to undergo this procedure."

Questions for the Committee:

- The evidence provided by the developer is updated and is directionally the same as that presented for the previous NQF review. Does the Committee agree there is no need to re-vote on Evidence?
- What is the relationship of this measure to patient outcomes?

- How strong is the evidence for this relationship?
- What can providers do to achieve a change in measure results?

Guidance from the Evidence Algorithm

Process measure based on systematic review (Box 3) à QQC presented (Box 4) à Quantity: high; Quality: high; Consistency: high (Box 5) à High (Box 5a) à High

The highest possible rating is high.

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.

From the 2016 Review:

- The developers provided data from a systematic review, which they stated demonstrated considerable variability in episiotomy rates.
- The developers also provided their own (National Perinatal Information Center) CY 2010 data. They stated that rates varied across centers, between 4.3% to 34.6%, with an average overall incidence of 16.2%.
- By 2014, the developer reports that overall incidence dropped from 11.5% to 7.2% and contends that this indicated a continued opportunity for improvement.

For the 2020 submission:

- The developer reported updated data for CY 2019:
    - Average across hospitals = 4.7%
    - Range: 0.0% to 13.9%

Disparities

Based on CY 2019 data:

- Rate by Race

Race	Rate (%)
Asian	7.8%
Black	2.7%
Native American/AK Native	3.3%
Native Hawaiian/Pacific Islander	4.0%
White	4.9%
Other	4.2%
Unknown	4.1%

- Rate by age group

Age Group	Rate (%)
less than 17 years of age	5.1%
17-24 years of age	4.6%
25-34 years of age	4.9%

35-44 years of age	4.4%
45+ years of age	3.8%
Total	4.7%

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)

1a. Evidence to Support Measure Focus: For all measures (structure, process, outcome, patient-reported structure/process), empirical data are required. How does the evidence relate to the specific structure, process, or outcome being measured? Does it apply directly or is it tangential? How does the structure, process, or outcome relate to desired outcomes? For maintenance measures–are you aware of any new studies/information that changes the evidence base for this measure that has not been cited in the submission? For measures derived from a patient report: Measures derived from a patient report must demonstrate that the target population values the measured outcome, process, or structure.

- No, I am not aware of any new evidence.

1b. Performance Gap: Was current performance data on the measure provided? How does it demonstrate a gap in care (variability or overall less than optimal performance) to warrant a national performance measure? Disparities: Was data on the measure by population subgroups provided? How does it demonstrate disparities in the care?

- Rates across centers/hospitals, as well as by race, demonstrate disparities in care.

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability: Specifications and Testing

2b. Validity: Testing; Exclusions; Risk-Adjustment; Meaningful Differences; Comparability; Missing Data

Reliability

2a1. Specifications requires the measure, as specified, to produce consistent (reliable) and credible (valid) results about the quality of care when implemented. For maintenance measures – no change in emphasis – specifications should be evaluated the same as with new measures.

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 – less emphasis if no new testing data provided.

Validity

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 – less emphasis if no new testing data provided.

2b2-2b6. Potential threats to validity should be assessed/addressed.

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.

Complex measure evaluated by Scientific Methods Panel? ☐ **Yes** ☒ **No**

Evaluators: NQF Staff

Reliability

- Reliability testing was performed at the measure score and data element levels.
    - The developer provided signal-to-noise reliability statistics to test the measure score (Mean: 4.8%; Standard Deviation: 3.1%; Standard Error: 0.32%; IQR of 4.4%)
    - In the validity section, the developer provided a Cohen's Kappa statistic and Inter-rater agreement to determine percent agreement between the encounters in each documentation method and test data element reliability (Kappa: 0.958; IRR: 97.7%).

Validity

- Validity testing was performed at the data element level
    - The developer provided several tests of validity (Sensitivity = 0.9725; Specificity = 0.9858; Positive Predicted Value (PPV) = 97.21%; Negative Predicted Value (NPV) = 98.60%)

Questions for the Committee regarding reliability:

- Do you have any concerns that the measure can be consistently implemented (i.e., are measure specifications adequate)?

Questions for the Committee regarding validity:

- Do you have any concerns regarding the testing used to illustrate the validity of this measure?

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

Scientific Acceptability: Preliminary Analysis Form

Measure Number: 0470

Measure Title: Incidence of Episiotomy

Type of measure:

☒ **Process** ☐ **Process: Appropriate Use** ☐ **Structure** ☐ **Efficiency** ☐ **Cost/Resource Use**
☐ **Outcome** ☐ **Outcome: PRO-PM** ☐ **Outcome: Intermediate Clinical Outcome** ☐ **Composite**

Data Source:

☒ **Claims** ☐ **Electronic Health Data** ☒ **Electronic Health Records** ☐ **Management Data**
☐ **Assessment Data** ☒ **Paper Medical Records** ☐ **Instrument-Based Data** ☐ **Registry Data**
☐ **Enrollment Data** ☐ **Other**

Level of Analysis:

☐ **Clinician: Group/Practice** ☐ **Clinician: Individual** ☒ **Facility** ☐ **Health Plan**
☐ **Population: Community, County or City** ☐ **Population: Regional and State**

☐ **Integrated Delivery System** ☐ **Other**

Measure is:

☐ **New** ☒ **Previously endorsed (**NOTE: Empirical validity testing is expected at time of maintenance review; if not possible, justification is required.)

RELIABILITY: SPECIFICATIONS

1. **Are submitted specifications precise, unambiguous, and complete so that they can be consistently implemented?** ☒ **Yes** ☐ **No**

   Submission document: "MIF_0480" document, items S.1-S.22

   ***NOTE**: NQF staff will conduct a separate, more technical, check of eCQM specifications, value sets, logic, and feasibility, so no need to consider these in your evaluation.*

2. **Briefly summarize any concerns about the measure specifications.**
   - No concerns

RELIABILITY: TESTING

Submission document: "MIF_0480" document for specifications, testing attachment questions 1.1-1.4 and section 2a2

3. **Reliability testing level** ☒ **Measure score** ☒ **Data element** ☐ **Neither**
4. **Reliability testing was conducted with the data source and level of analysis indicated for this measure** ☒ **Yes** ☐ **No**
5. If score-level and/or data element reliability testing was NOT conducted or if the methods used were NOT appropriate, was **empirical VALIDITY testing** of patient-level data conducted?

   ☐ **Yes** ☐ **No**

6. **Assess the method(s) used for reliability testing**

   Submission document: Testing attachment, section 2a2.2
   - The developer provided score-level reliability for the measure using the beta-binomial model (signal to noise). Testing was performed on a dataset of 215,912 vaginal delivery inpatient encounters across a set of 14 hospitals.
   - The developer reported that this subset of hospitals are all part of the Council of Women's and Infants' Specialty Hospitals (CWISH) and that this is a nationally representative subgroup of hospitals

7. **Assess the results of reliability testing**

   Submission document: Testing attachment, section 2a2.3
   - For the 2020 submission, the developer reported the following signal to noise reliability statistics:
     - Mean: 0.9677
     - Standard Deviation: 0.0380
     - Minimum: 0.8220
     - 25th-75th Percentile: 0.9646-0.9893
     - IQR = 0.0247
   - The developer noted that average score 0.9677 is acceptable reliability for most of the hospitals.

8. Was the method described and appropriate for assessing the proportion of variability due to real differences among measured entities? NOTE: If multiple methods used, at least one must be appropriate.

   Submission document: Testing attachment, section 2a2.2 LINK

   ☒ **Yes**

☐ **No**

☐ **Not applicable** (score-level testing was not performed)

9. Was the method described and appropriate for assessing the reliability of ALL critical data elements?

   Submission document: Testing attachment, section 2a2.2LINK

   ☒ **Yes**

   ☐ **No**

   ☐ **Not applicable** (data element testing was not performed)

10. **OVERALL RATING OF RELIABILITY** (taking into account precision of specifications and all testing results):

    ☒ **High** (NOTE: Can be HIGH only if score-level testing has been conducted)

    ☐ **Moderate** (NOTE: Moderate is the highest eligible rating if score-level testing has not been conducted)

    ☐ **Low** (NOTE: Should rate LOW if you believe specifications are NOT precise, unambiguous, and complete or if testing methods/results are not adequate)

    ☐ **Insufficient** (NOTE: Should rate INSUFFICIENT if you believe you do not have the information you need to make a rating decision)

11. **Briefly explain rationale for the rating of OVERALL RATING OF RELIABILITY and any concerns you may have with the approach to demonstrating reliability.**
    - The developer provided a score-level reliability statistic of 0.9
    - The developer attests that the testing sample was nationally representative
    - Are submitted specifications precise, unambiguous, and complete so that they can be consistently implemented? (Box 1) → Was empirical reliability testing conducted using statistical tests with the measure as specified? (Box 2) → Was reliability testing conducted with computed performance scores for each measured entity? (Box 4) → Was the method described and appropriate for assessing the proportion of variability due to real differences among measured entities? (Box 5) → Is there a high certainty or confidence that the performance measure scores are reliable? (Box 6a) → High

VALIDITY: ASSESSMENT OF THREATS TO VALIDITY

12. **Please describe any concerns you have with measure exclusions.**

    Submission document: Testing attachment, section 2b2.

    - A code indicating shoulder dystocia would result in exclusion from the denominator. The developer provides a clinical rational for this exclusion.
    - The developer examined a coded and manually abstracted data set of encounters with qualifying denominator exclusions and reported that 2.2% of the sample was excluded.
    - The developer calculated a Cohen's Kappa statistic of 97.6% to determine percent agreement between the encounters in each documentation method. The developer reported this suggests accurate coding and reliable exclusion of these patients from the population.
    - No concerns.

13. **Please describe any concerns you have regarding the ability to identify meaningful differences in performance.**

    Submission document: Testing attachment, section 2b4

    - The developer presented the following descriptive statistics for each facility in its database:
      - Mean: 4.8%
      - Standard Deviation: 3.1%
      - Standard Error: 0.32%
      - IQR of 4.4%

- The developer stated that the hospitals that performed two standard deviations above the mean performed worse than the average hospital.
- The developer reports that out of 87 hospitals, three facilities had a performance rate greater than two standard deviations above the mean and therefore performed significantly worse than average.
- No concerns.

14. **Please describe any concerns you have regarding comparability of results if multiple data sources or methods are specified.**
Submission document: Testing attachment, section 2b5.
    - Not applicable
15. **Please describe any concerns you have regarding missing data.**
Submission document: Testing attachment, section 2b6.
    - Not applicable
16. **Risk Adjustment**

16a. **Risk-adjustment method** ☒ **None** ☐ **Statistical model** ☐ **Stratification**

16b. **If not risk-adjusted, is this supported by either a conceptual rationale or empirical analyses?**

☐ Yes ☒ No ☐ Not applicable

- The developer states that patient-level sociodemographic variables were examined, but determined that neither adjustment nor stratification were required "to achieve fair comparisons across measured entities."

16c. **Social risk adjustment:**

16c.1 Are social risk factors included in risk model? ☐ Yes ☐ No ☒ Not applicable

16c.2 Conceptual rationale for social risk factors included? ☐ Yes ☐ No

16c.3 Is there a conceptual relationship between potential social risk factor variables and the measure focus? ☐ Yes ☐ No

16d. **Risk adjustment summary:**

- Not applicable

16d.1 All of the risk-adjustment variables present at the start of care? ☐ Yes ☐ No

16d.2 If factors not present at the start of care, do you agree with the rationale provided for inclusion? ☐ Yes ☐ No

16d.3 Is the risk adjustment approach appropriately developed and assessed? ☐ Yes ☐ No

16d.4 Do analyses indicate acceptable results (e.g., acceptable discrimination and calibration) ☐ Yes ☐ No

16d.5. Appropriate risk-adjustment strategy included in the measure? ☐ Yes ☐ No

16e. **Assess the risk-adjustment approach**

- The measure is not risk adjusted. The developer attests that social risk factor information is not available, but the developer did not provide a conceptual rational for this approach nor did the developer indicate a literature review was performed.

For cost/resource use measures ONLY:

17. **Are the specifications in alignment with the stated measure intent?**

☐ **Yes** ☐ **Somewhat** ☐ **No (If "Somewhat" or "No", please explain)**

18. **Describe any concerns of threats to validity related to attribution, the costing approach, carve outs, or truncation (approach to outliers):**

VALIDITY: TESTING

19. **Validity testing level:** ☐ **Measure score** ☒ **Data element** ☐ **Both**
20. **Method of establishing validity of the measure score:**

    ☐ **Face validity**

    ☐ **Empirical validity testing of the measure score**

    ☒ **N/A (score-level testing not conducted)**
21. **Assess the method(s) for establishing validity**

    Submission document: Testing attachment, section 2b2.2.

    - In the nationally representative sample of hospitals described above, the developer examined all critical data elements for this measure. The developer reported that they excluded one hospital for a total of 13 because they did not respond in time to be included for analysis.
    - The developer examined a coded and manually abstracted data set of encounters with qualifying numerator and denominator exclusions and then calculated a Cohen's Kappa statistic to determine percent agreement between the encounters in each documentation method.
    - The developer also analyzed the sensitivity and specificity of the dataset and positive and negative predictive values to test the accuracy of the measure.
22. **Assess the results(s) for establishing validity**

    Submission document: Testing attachment, section 2b2.3.

    - Sensitivity = 0.9725
    - Specificity = 0.9858
    - Positive Predicted Value (PPV) = 97.21%
    - Negative Predicted Value (NPV) = 98.60%
23. **Was the method described and appropriate for assessing conceptually and theoretically sound hypothesized relationships?**

    Submission document: Testing attachment, section 2b1.

    ☒ **Yes**

    ☐ **No**

    ☐ **Not applicable** (score-level testing was not performed)
24. **Was the method described and appropriate for assessing the accuracy of ALL critical data elements?**

    NOTE that data element validation from the literature is acceptable.

    Submission document: *Testing attachment, section 2b1.*

    ☒ **Yes**

    ☐ **No**

    ☐ **Not applicable** (data element testing was not performed)
25. **OVERALL RATING OF VALIDITY taking into account the results and scope of all testing and analysis of potential threats.**

    ☐ **High** (NOTE: Can be HIGH only if score-level testing has been conducted)

    ☒ **Moderate** (NOTE: Moderate is the highest eligible rating if score-level testing has NOT been conducted)

    ☐ **Low** (NOTE: Should rate LOW if you believe that there are threats to validity and/or relevant threats to validity were not assessed OR if testing methods/results are not adequate)

☐ **Insufficient** (NOTE: For instrument-based measures and some composite measures, testing at both the score level and the data element level is required; if not conducted, should rate as INSUFFICIENT.)

26. **Briefly explain rationale for rating of OVERALL RATING OF VALIDITY and any concerns you may have with the developers' approach to demonstrating validity.**
    - The developer tested all critical data elements.
    - Were all potential threats to validity that are relevant to the measure empirically assessed? (box 1) → was empirical validity testing conducted using the measure as specified and appropriate statistical test? (box 2) → was validity testing conducted with computed performance measure scores for each measured entity? (box 5) → was validity testing conducted with patient level data elements? (box 9) → was the method described and appropriate for assessing the accuracy of ALL critical data elements? (box 10) → is there a high or moderate certainty or confidence that the data used in the measure are valid? (box 11a) → Moderate

ADDITIONAL RECOMMENDATIONS

27. **If you have listed any concerns in this form, do you believe these concerns warrant further discussion by the multi-stakeholder Standing Committee? If so, please list those concerns below.**

Committee Pre-evaluation Comments:

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

2a1. Reliability-Specifications: Which data elements, if any, are not clearly defined? Which codes with descriptors, if any, are not provided? Which steps, if any, in the logic or calculation algorithm or other specifications (e.g., risk/case-mix adjustment, survey/sampling instructions) are not clear? What concerns do you have about the likelihood that this measure can be consistently implemented?

- No concerns

2a2. Reliability - Testing: Do you have any concerns about the reliability of the measure?

The Split Sample reliability correlation was acceptable at 0.632; I was a little concerned about the range of Signal-to-Noise results from 0.34 to 0.99 (median 0.79)

- The average signal-to-noise score of 0.9677 seems high. While the developer noted that the score is acceptable reliability for most hospitals, this explanation does not fully address concerns.

2b1. Validity -Testing: Do you have any concems with the testing results?

- No concerns.

2b2-3. Other Threats to Validity (Exclusions, Risk Adjustment) 2b2. Exclusions: Are the exclusions consistent with the evidence? Are any patients or patient groups inappropriately excluded from the measure? 2b3. Risk Adjustment: If outcome (intermediate, health, or PRO-based) or resource use performance measure: Is there a conceptual relationship between potential social risk factor variables and the measure focus? How well do social risk factor variables that were available and analyzed align with the conceptual description provided? Are all of the risk-adjustment variables present at the start of care (if not, do you agree with the rationale provided)? Was the risk adjustment (case-mix adjustment) appropriately developed and tested? Do analyses indicate acceptable results? Is an appropriate risk-adjustment strategy included in the measure?

- The developer does not provide a sufficient rationale for its risk adjustment approach. It seems to me like social risk factor information should be available and assessed.

2b4-7. Threats to Validity (Statistically Significant Differences, Multiple Data Sources, Missing Data) 2b4. Meaningful Differences: How do analyses indicate this measure identifies meaningful differences about quality? 2b5. Comparability of performance scores: If multiple sets of specifications: Do analyses indicate they produce comparable results? 2b6. Missing data/no response: Does missing data constitute a threat to the validity of this measure?

- This measure appears to identify meaningful differences about quality of care.

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 reported that data are generated by and used by healthcare personnel during the provision of care, coded by someone other than person obtaining original information (e.g., DRG, ICD-10 data) and that all data elements are in defined fields in electronic sources.
    - The developer reports that the measure is calculated using MS-DRG and ICD-10 code criteria

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

3. **Feasibility: Which of the required data elements are not routinely generated and used during care delivery? Which of the required data elements are not available in electronic form (e.g., EHR or other electronic sources)? What are your concerns about how the data collection strategy can be put into operational use?**
    - 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

4a. Use (4a1. Accountability and Transparency; 4a2. Feedback on measure)

4a. 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.

4a.1. 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.

Current uses of the measure

Publicly reported? ☒ **Yes** ☐ **No**

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

Accountability program details

- The developer reported the measure is publicly reported and used for accountability as part of Leapfrog Group and the National Perinatal Information Center, Inc. (NPIC) Metric.

4a.2. Feedback on the measure by those being measured or others. Three criteria demonstrate feedback: 1) those being measured have been given performance results or data, as well as assistance with interpreting the measure results and data; 2) those being measured and other users have been given an opportunity to provide feedback on the measure performance or implementation; 3) this feedback has been considered when changes are incorporated into the measure

Feedback on the measure by those being measured or others

- The developer reported:
    - Quarterly webinars are offered to hospitals to disseminate performance results
    - Measure users also receive data and are assisted with interpreting that data.
    - Measure users are able to reach out for assistance with improving performance rates
    - Measure users are satisfied with the measure and have not reported feedback warranting significant change to the measure

Additional Feedback:

- Not Reviewed by Measure Applications Partnership.

***Questions for the Committee**:*

- Can the performance results be used to further the goal of high-quality, efficient healthcare?

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

4b. Usability (4a1. Improvement; 4a2. Benefits of measure)

4b. Usability 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.

4b.1 Improvement. Progress toward achieving the goal of high-quality, efficient healthcare for individuals or populations is demonstrated.

Improvement results

- The performance trend for this measure is as follows:
    - CY 2010: 11.5%
    - CY 2014: 7.2%
    - CY 2019: 4.7%

4b2. Benefits vs. harms. 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).

Unexpected findings (positive or negative) during implementation

- The developer reported one unintended consequence related to ICD-9 coding and the actions taken to mitigate this issue:
    - Measure users indicated issues with coding of lacerations resulting from an episiotomy tearing and an update was made to include episiotomy and repair of laceration procedures. The developer reports that this update reduced inaccuracies in reporting, given their reported high degree of match between administrative and abstracted data.

Potential harms

- None identified by the developer

Additional Feedback:

- None reported

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

4a1. Use - Accountability and Transparency: How is the measure being publicly reported? Are the performance results disclosed and available outside of the organizations or practices whose performance is measured? For maintenance measures - which accountability applications is the measure being used for? For new measures - if not in use at the time of initial endorsement, is a credible plan for implementation provided? 4a2. Use - Feedback on the measure: Have those being measured been given performance results or data, as well as assistance with interpreting the measure results and data? Have those being measured or other users been given an opportunity to provide feedback on the measure performance or implementation? Has this feedback has been considered when changes are incorporated into the measure?

- No concerns with the developer's strategy for disseminating measure data to and soliciting feedback from end users.

4b1. Usability – Improvement: How can the performance results be used to further the goal of high-quality, efficient healthcare? If not in use for performance improvement at the time of initial endorsement, is a credible rationale provided that describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations? 4b2. Usability – Benefits vs. harms: Describe any actual unintended consequences and note how you think the benefits of the measure outweigh them.

- I am not concerned about any unintended consequences from collecting measure data.

Criterion 5: Related and Competing Measures

Related or competing measures

- None

Harmonization

- Not applicable

Committee Pre-evaluation Comments: Criterion 5:
Related and Competing Measures

5. Related and Competing: Are there any related and competing measures? If so, are any specifications that are not harmonized? Are there any additional steps needed for the measures to be harmonized?

- No--not applicable.

Public and Member Comments

Comments and Member Support/Non-Support Submitted as of: 01/21/2021

- **No NQF Members have submitted support/non-support choices as of this date.**
- **No Public or NQF Member comments submitted as of this date.**

1. Evidence and Performance Gap – 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 sub criteria 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_attachment_2020-11-09-637412820352568757.docx

1a.1 For Maintenance of Endorsement: Is there new evidence about the measure since the last update/submission?

Do not remove any existing information. If there have been any changes to evidence, the Committee will consider the new evidence. Please use the most current version of the evidence attachment (v7.1). Please use red font to indicate updated evidence.

Yes

1a. Evidence (subcriterion 1a)

Measure Number (*if previously endorsed*)**:** 0470

Measure Title: Incidence of Episiotomy

IF the measure is a component in a composite performance measure, provide the title of the Composite Measure here:

Date of Submission: 11/9/2020

1a.1.This is a measure of: (*should be consistent with type of measure entered in De.1*)

Outcome

☐ Outcome:

☐Patient-reported outcome (PRO):

PROs include HRQoL/functional status, symptom/symptom burden, experience with care, health-related behaviors. (*A PRO-based performance measure is not a survey instrument. Data may be collected using a survey instrument to construct a PRO measure.)*

☐ Intermediate clinical outcome (*e.g., lab value*):

☒ Process: **Incidence of Episiotomy**

☐ Appropriate use measure:

☐ Structure:

☐ Composite:

1a.2 LOGIC MODEL Diagram or briefly describe the steps between the healthcare structures and processes (e.g., interventions, or services) and the patient's health outcome(s). The relationships in the diagram should be easily understood by general, non-technical audiences. Indicate the structure, process or outcome being measured.

During a vaginal birth an episiotomy may be considered if the baby needs to be delivered expeditiously, i.e., shoulder dystocia (baby stuck behind the pelvic bone) and an operative delivery (forceps or vacuum) is required. An incision (midline or mediolateral) is made in the perineum. After delivery the incision is repaired. For years routine episiotomy was thought to prevent more extensive vaginal tears during childbirth and allow for better healing. It was also thought to preserve the pelvic floor musculature. Research has since shown this is not the case and might be detrimental.

1a.3 Value and Meaningfulness: IF this measure is derived from patient report, provide evidence that the target population values the measured ***outcome, process, or structure*** and finds it meaningful. (Describe how and from whom their input was obtained.)

NA

****RESPOND TO ONLY ONE SECTION BELOW -EITHER 1a.2, 1a.3 or 1a.4) ****

1a.2 FOR OUTCOME MEASURES including PATIENT REPORTED OUTCOMES - Provide empirical data demonstrating the relationship between the outcome (or PRO) to at least one healthcare structure, process, intervention, or service.

1a.3. SYSTEMATIC REVIEW(SR) OF THE EVIDENCE (for INTERMEDIATE OUTCOME, PROCESS, OR STRUCTURE PERFORMANCE MEASURES, INCLUDING THOSE THAT ARE INSTRUMENT-BASED) If the evidence is not based on a systematic review go to section 1a.4) If you wish to include more than one systematic review, add additional tables.

What is the source of the systematic review of the body of evidence that supports the performance measure? A systematic review is a scientific investigation that focuses on a specific question and uses explicit, prespecified scientific methods to identify, select, assess, and summarize the findings of similar but separate studies. It may include a quantitative synthesis (meta-analysis), depending on the available data. (IOM)

☒ Clinical Practice Guideline recommendation (with evidence review)

☐ US Preventive Services Task Force Recommendation

☐ Other systematic review and grading of the body of evidence (*e.g., Cochrane Collaboration, AHRQ Evidence Practice Center*)

☐ Other

Systematic Review	Evidence
Source of Systematic Review: • Title • Author • Date • Citation, including page number • URL	ACOG Practice Bulletin no. 71 April 2006 Practice Bulletin No. 165: Prevention and Management of Obstetric Lacerations at Vaginal Delivery, Obstetrics & Gynecology: July 2016 - Volume 128 - Issue 1 - p e1-e15 doi: 10.1097/AOG.0000000000001523
Quote the guideline or recommendation verbatim about the process, structure or intermediate outcome being measured. If not a guideline, summarize the conclusions from the SR.	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.
Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	Entity that graded evidence: ACOG System Used for Grading Body of Evidence: GRADE Grade Assigned to Body of Evidence: A
Provide all other grades and definitions from the evidence grading system	*
Grade assigned to the **recommendation** with definition of the grade	Entity that graded evidence: ACOG System Used for Grading Body of Evidence: GRADE Grade Assigned to Body of Evidence: A
Provide all other grades and definitions from the recommendation grading system	*
Body of evidence: • Quantity – how many studies? • Quality – what type of studies?	Quantity: A PubMed search, reveals 2160 articles on episiotomy of which 195 are reviews. 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.

Systematic Review	Evidence
Estimates of benefit and consistency across studies	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.
What harms were identified?	*
Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	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 https://www.cdc.gov/nchs/data/nhds/4procedures/2009pro4_numberrate.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)

Systematic Review	Evidence
	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 follow-up of alternative methods used in perineal repair. BJOG 2001;108:34–40. (Level II-2)

Systematic Review	Evidence
	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)

Systematic Review	Evidence
	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)

Systematic Review	Evidence
	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)

Systematic Review	Evidence
	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) Papadakis K., Myriknas S. Standardizing indications for episiotomy: A narrative review of contemporary clinical evidence. Journal of Pelvic, Obstetric and Gynaecological Physiotherapy. 2020 (126) (pp 5-10), 2020. Date of Publication: Spring 2020. Sultan AH , Thakar R , Ismail KM , Kalis V, Laine K, Räisänen SH, de Leeuw VW.. The role of mediolateral episiotomy during operative vaginal delivery. Eur J Obstet Gynecol Reprod Biol. 2019 Sep;240:192-196. doi: 10.1016/j.ejogrb.2019.07.005. Epub 2019 Jul 9. Frigerio M, Mastrolia SA , Spelzini F , Manodoro S, Yohay D, Weintraub AY. Long-term effects of episiotomy on urinary incontinence and pelvic organ prolapse: a systematic review. Arch Gynecol Obstet. 2019 Feb;299(2):317-325. doi: 10.1007/s00404-018-5009-9. Epub 2018 Dec 18. Christophe Clesse, Joëlle Lighezzolo-Alnot, Sylvie De Lavergne, Sandrine Hamlin & Michèle Scheffler (2018) Statistical trends of episiotomy around the world: Comparative systematic review of changing practices, Health Care for Women International, 39:6, 644-662, DOI: 10.1080/07399332.2018.1445253 Hong Jiang, Xu Qian, Guillermo Carroli, Paul Garner. Selective versus routine use of episiotomy for vaginal birth. Cochrane Database Syst Rev. 2017 Feb 8;2(2):CD000081. doi: 10.1002/14651858.CD000081.pub3 Christophe Clesse, Joëlle Lighezzolo-Alnot, Sylvie De Lavergne, Sandrine Hamlin & Michèle Scheffler (2019) Socio-historical evolution of the episiotomy practice: A literature review, Women & Health, 59:7, 760-774, DOI: 10.1080/03630242.2018.1553814 Christophe Clesse, Joëlle Lighezzolo-Alnot, Sylvie De Lavergne, Sandrine Hamlin & Michèle Scheffler (2019) Factors related to episiotomy practice: an evidence-based medicine systematic review, Journal of Obstetrics and Gynaecology, 39:6, 737-747, DOI: 10.1080/01443615.2019.1581741

Systematic Review	Evidence
	Pereira, G.M.V., Hosoume, R.S., de Castro Monteiro, M.V. et al. Selective episiotomy versus no episiotomy for severe perineal trauma: a systematic review with meta-analysis. Int Urogynecol J 31, 2291–2299 (2020). https://doi.org/10.1007/s00192-020-04308-2

*cell intentionally left blank

_____

1a.4 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.4.1 Briefly SYNTHESIZE the evidence that supports the measure. A list of references without a summary is not acceptable.

1a.4.2 What process was used to identify the evidence?

1a.4.3. Provide the citation(s) for the evidence.

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., how the measure will improve the quality of care, the benefits or improvements in quality envisioned by use of this measure)*

If a COMPOSITE (e.g., combination of component measure scores, all-or-none, any-or-none), SKIP this question and answer the composite questions.

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 fasciitis. 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 maintenance of endorsement. 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 include.) This information also will be used to address the sub-criterion on improvement (4b1) 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 public 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 maintenance of endorsement. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included.) For measures that show high levels of performance, i.e., "topped out", disparities data may demonstrate an opportunity for improvement/gap in care for certain sub-populations. This information also will be used to address the sub-criterion on improvement (4b1) 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.

CY 2019 update:

average across hospitals = 4.7%

range: low if 0.0% to a high of 13.9%

Rate by age group:

Age Group	rate
less than 17 years of age	5.1%
17-24 years of age	4.6%
25-34 years of age	4.9%
35-44 years of age	4.4%
45+ years of age	3.8%
Total	4.7%

Rate by Race:

Race	rate
Asian	7.8%
Black	2.7%
Native American/AK Native	3.3%
Native Hawaiian/Pacific Islander	4.0%
White	4.9%
Other	4.2%
Unknown	4.1%

1b.5. If no or limited data on disparities from the measure as specified is reported in 1b.4, 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

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 sub criteria 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 Health

De.6. Non-Condition Specific*(check all the areas that apply):*

Safety : Overuse

De.7. Target Population Category *(Check all the populations for which the measure is specified and tested if any):*

Women

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/data-partnership/nqf-measure-steward/

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**:** ICD-10_Codes_NQF_Episiotomy_FINAL_NQF_Submission-636826277172046258-637139158102535322.xlsx

S.2c. Is this an instrument-based measure (i.e., data collected via instruments, surveys, tools, questionnaires, scales, etc.)? Attach copy of instrument if available.

No, this is not an instrument-based measure **Attachment:**

S.2d. Is this an instrument-based measure (i.e., data collected via instruments, surveys, tools, questionnaires, scales, etc.)? Attach copy of instrument if available.

S.3.1. For maintenance of endorsement: Are there changes to the specifications since the last updates/submission. If yes, update the specifications for S1-2 and S4-22 and explain reasons for the changes in S3.2.

No

S.3.2. For maintenance of endorsement, please briefly describe any important changes to the measure specifications since last measure update and explain the reasons.

No changes except to update for use of ICD 10 codes beginning 10/1/2015 and the MS DRGs for vaginal delivery starting with 10/2018 discharges

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) DO NOT include the rationale for the measure.*

IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm (S.14).

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. 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, time period for data collection, 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 (S.14).

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.6. 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.7. Denominator Details *(All information required to identify and calculate the target population/denominator such as definitions, time period for data collection, 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 target population is identified. Calculation of the risk-adjusted outcome should be described in the calculation algorithm (S.14).

Any woman with a vaginal delivery calculated by either MS DRG 774, 775, 767, 768: MS DRGs starting with 10/1/2018 discharges: 768, 796, 797, 798, 805 and 807

S.8. 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.9. Denominator Exclusion Details *(All information required to identify and calculate exclusions from the denominator such as definitions, time period for data collection, 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.10. Stratification Information *(Provide all information required to stratify the measure results, if necessary, including the stratification variables, definitions, specific data collection items/responses, code/value sets, and the risk-model covariates and coefficients for the clinically-adjusted version of the measure when appropriate – 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.11. Risk Adjustment Type (Select type. Provide specifications for risk stratification in measure testing attachment)

No risk adjustment or risk stratification

If other:

S.12. Type of score:

Rate/proportion

If other:

S.13. 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.14. Calculation Algorithm/Measure Logic (*Diagram or 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; time period for data, 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.15. Sampling *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*

IF an instrument-based performance measure (e.g., PRO-PM), identify whether (and how) proxy responses are allowed.

NA

S.16. Survey/Patient-reported data *(If measure is based on a survey or instrument, provide instructions for data collection and guidance on minimum response rate.)*

Specify calculation of response rates to be reported with performance measure results.

S.17. Data Source *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*

If other, please describe in S.18.

Claims, Electronic Health Data, Electronic Health Records, Paper Medical Records

S.18. Data Source or Collection Instrument *(Identify the specific data source/data collection instrument (e.g. name of database, clinical registry, collection instrument, etc., and describe how data are collected.)*

IF instrument-based, identify the specific instrument(s) and standard methods, modes, and languages of administration.

UB04 claims data.

S.19. 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.20. Level of Analysis *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*

Facility

S.21. Care Setting *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*

Inpatient/Hospital

If other:

S.22. 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.)*

2. Validity – See attached Measure Testing Submission Form

NQF_testing_0470_submit-637378714863816283.docx

2.1 For maintenance of endorsement

Reliability testing: If testing of reliability of the measure score was not presented in prior submission(s), has reliability testing of the measure score been conducted? If yes, please provide results in the Testing attachment. Please use the most current version of the testing attachment (v7.1). Include information on all testing conducted (prior testing as well as any new testing); use red font to indicate updated testing.

Yes

2.2 For maintenance of endorsement

Has additional empirical validity testing of the measure score been conducted? If yes, please provide results in the Testing attachment. Please use the most current version of the testing attachment (v7.1). Include information on all testing conducted (prior testing as well as any new testing); use red font to indicate updated testing.

Yes

2.3 For maintenance of endorsement

Risk adjustment: For outcome, resource use, cost, and some process measures, risk-adjustment that includes social risk factors is not prohibited at present. Please update sections 1.8, 2a2, 2b1,2b4.3 and 2b5 in the Testing attachment and S.140 and S.11 in the online submission form. NOTE: These sections must be updated even if social risk factors are not included in the risk-adjustment strategy. You MUST use the most current version of the Testing Attachment (v7.1) -- older versions of the form will not have all required questions.

No - This measure is not risk-adjusted

Measure Testing (subcriteria 2a2, 2b1-2b6)

Measure Number (*if previously endorsed*)**:** 0470
Measure Title: Incidence of Episiotomy
Date of Submission: 10/9/2020

Type of Measure:

Measure	Measure (continued)
☐ **Outcome (*including PRO-PM*)**	☐ **Composite – *STOP – use composite testing form***
☐ Intermediate Clinical Outcome	☐ Cost/resource
☒ Process *(including Appropriate Use)*	☐ Efficiency
☐ Structure	*

*cell intentionally left blank

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.17***)	**Measure Tested with Data From:**
☒ abstracted from paper record	☒ abstracted from paper record
☒ claims	☒ claims
☐ registry	☐ registry
☒ abstracted from electronic health record	☒ abstracted from electronic health record
☐ eMeasure (HQMF) implemented in EHRs	☐ eMeasure (HQMF) implemented in EHRs
☐ other:	☐ other:

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*).

National Perinatal Information Center (NPIC) Perinatal Center Database (PCDB) comprising of Administrative (UB-04) data elements as well as supplemental perinatal variables. All database records are validated prior to inclusion. For this measure 215,912 vaginal delivery inpatient encounters were queried for the database from a recent twelve-month period.

http://www.npic.org/perinatal-center-database/

1.3. What are the dates of the data used in testing? **4/1/2018 – 3/31/2019**

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.20***)	**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*)

For validity and measure-level reliability the full PCDB was used for testing. For data element-level reliability a subset of 14 hospitals comprised of the Council of Women's and Infants' Specialty Hospitals (CWISH) was included. CWISH is a membership organization of non-profit hospitals that are leaders in providing services to women and infants. It is a unique, nationally representative subgroup of National Perinatal Information Center (NPIC) hospitals with large maternity services.

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 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 totaled 66,306 eligible deliveries and 9,626 episiotomy cases, Period 2: 31,496 eligible deliveries and 4,259 episiotomy cases.

The testing was performed on the full nationally representative PCDB dataset of 215,912 vaginal delivery inpatient encounters for the 4/1/2018 – 3/31/2019 time period. Below are two tables showing distribution by Age Group and Race:

Age Group	n	%
less than 17 years of age	1,168	0.5%
17-24 years of age	45,847	21.2%
25-34 years of age	127,225	58.9%
35-44 years of age	41,325	19.1%
45+ years of age	347	0.2%
Total	**215,912**	**100.0%**
Race	**n**	**%**
Asian	13,974	6.5%
Black	42,003	19.5%
Native American/AK Native	1,196	0.6%
Native Hawaiian/Pac Islander	817	0.4%
White	116,887	54.1%
Other	12,520	5.8%
Unknown	28,515	13.2%
Total	**215,912**	**100.0%**

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.

For validity and measure-level reliability the full PCDB was used for testing. For data element-level reliability a subset of 14 CWISH hospitals within the PCDB was used. This subset represents 36% of the delivery volume in the PCDB for the 4/1/2018 – 3/31/2019 time period.

1.8 What were the social risk factors that were available and analyzed? For example, patient-reported data (e.g., income, education, language), proxy variables when social risk data are not collected from each patient (e.g. census tract), or patient community characteristics (e.g. percent vacant housing, crime rate) which do not have to be a proxy for patient-level data.

No social risk factors were available 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*)

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 totaled 66,306 eligible deliveries and 9,626 episiotomy cases, Period 2: 31,496 eligible deliveries and 4,259 episiotomy cases.

Per NQF guidelines, for Data Element testing see validity testing section below.

For measure score reliability testing the numerators, denominators and rates were calculated from the PCDB at the hospital level. A beta-binomial model was used to calculate reliability scores for each hospital. This method was chosen after review of the "The Reliability of Provider Profiling: A Tutorial[1]" paper by John L. Adams which states "Fundamentally, reliability is the measure of whether you can tell one physician, from another." In this case we look at the hospital/facility-level, and test whether our measure can reliably distinguish performance differences between two entities.

References

Adams, John L. The Reliability of Provider Profiling: a Tutorial. Santa Monica, CA: RAND Corporation. 2009. Source: http://www.rand.org/pubs/technical_reports/TR653

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*)

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.

Measure score testing results (Reliability) were calculated at the hospital level. Summary statistics of the beta-binomial model reliability scores:

Statistic	Reliability
Mean	0.9677
Standard Deviation	0.0380
Minimum	0.8220
25th Percentile	0.9646
50th Percentile	0.9801
75th Percentile	0.9893
Maximum	1.0000
Interquartile Range	0.0247

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 mean reliability score of 0.9677 is above accepted minimum thresholds for reliability. This measure can reliably determine differences between entities measured.

2b1. VALIDITY TESTING

2b1.1. 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*) **NOTE**: Empirical validity testing is expected at time of maintenance review; if not possible, justification is required.

2b1.2. 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)*

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 totaled 66,306 eligible deliveries and 9,626 episiotomy cases, Period 2: 31,496 eligible deliveries and 4,259 episiotomy cases.

In addition to auditing their sample of cases, hospitals were asked to rate the reliability, validity, feasibility and usability of this measure

Using a sample of 14 hospitals from the PCDB a sample of 5% (or no less than 15) of the qualifying numerator encounters (vaginal delivery encounters where an episiotomy was coded) and additional denominator inclusions (vaginal delivery encounters where an episiotomy was NOT coded) were sent to each respective hospital (N = 645). These encounters were manually reviewed and results indicating whether the encounters were accurately grouped were sent back. From this validated dataset an inter rater reliability was calculated comparing the coded calculation of the measure through the PCDB compared to the manual abstraction. A Cohen's Kappa coefficient was calculated, in addition to measuring simple percent agreement between the two methods, in order to control for agreement occurring by chance. We also included Sensitivity and Specificity results from the analysis.

2b1.3. What were the statistical results from validity testing? (*e.g., correlation; t-test*)
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 the provider level since providers will determine whether to perform an episiotomy or not

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.

For the numerator and qualifying denominator dataset (Episiotomy or NO Episiotomy, both with no diagnosis of shoulder dystocia) there were 430 encounters tested from 13 responding of the 14 hospital dataset. Note that 13 hospitals responded with manual verification of the data, one hospital did not respond in the necessary timeframe and were excluded. The percent agreement between coded calculation and manual abstraction was 97.9% with a Kappa coefficient of 0.958. Sensitivity was 0.9725 and Specificity was 0.9858. Below is a confusion matrix table showing the results:

Confusion Matrix

coded	manual abstraction: No Episiotomy	manual abstraction: Episiotomy
No Episiotomy	212	3
Episiotomy	6	209

Positive Predicted Value (PPV) = 97.21%
Negative Predicted Value (NPV) = 98.60%

2b1.4. 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 agreement level of 97.9% and Kappa coefficient of 0.958 both indicate that the metric definition accurately calculates a hospital's episiotomy rate and is reliable. The Sensitivity and Specificity results show that Episiotomy procedure is coded accurately and encounters are classified correctly.

_____

2b2. EXCLUSIONS ANALYSIS

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

2b2.1. 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*)

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%

Using a sample of 14 hospitals from the PCDB a sample of 5% (or no less than 15) of the exclusion encounters (vaginal delivery where shoulder dystocia was coded) were sent to each respective hospital. These encounters were manually reviewed and results indicating whether the encounters were accurately grouped were sent back. From this validated dataset an inter rater reliability was calculated comparing the coded calculation of the measure through the PCDB compared to the manual abstraction. A simple agreement calculation was calculated. Of the 215,912 encounters in the dataset 4,898 (2.2%) were coded with shoulder dystocia.

2b2.2. 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*)

For the 215 encounter exclusion dataset (coded with shoulder dystocia) the percent agreement was 97.6%.

2b2.3. 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 high percent agreement of 97.6% shows that shoulder dystocia cases are accurately coded and reliably excluded from the measure population.

_____

2b3. 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.

2b3.1. What method of controlling for differences in case mix is used?

☒ **No risk adjustment or stratification**

☐ **Statistical risk model with risk factors**

☐ **Stratification by risk categories**

☐ **Other,**

2b3.1.1 If using a statistical risk model, provide detailed risk model specifications, including the risk model method, risk factors, coefficients, equations, codes with descriptors, and definitions.

2b3.2. 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.

Exploratory analyses were performed on the PCDB dataset to review any noticeable or significant differences in outcomes among various patient characteristics. While minor variations were seen among patient populations it was determined risk adjustment or stratification was not needed to achieve fair comparisons across measured entities.

2b3.3a. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors (clinical factors or social risk 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*) **Also discuss any "ordering" of risk factor inclusion**; for example, are social risk factors added after all clinical factors?

2b3.3b. How was the conceptual model of how social risk impacts this outcome developed? Please check all that apply:

☐ **Published literature**

☐ **Internal data analysis**

☐ **Other (please describe)**

2b3.4a. What were the statistical results of the analyses used to select risk factors?

2b3.4b. Describe the analyses and interpretation resulting in the decision to select social risk 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.)* **Also describe the impact of adjusting for social risk (or not) on providers at high or low extremes of risk.**

2b3.5. 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 2b3.9

2b3.6. Statistical Risk Model Discrimination Statistics (*e.g., c-statistic, R-squared*)**:**

2b3.7. Statistical Risk Model Calibration Statistics (*e.g., Hosmer-Lemeshow statistic*):

2b3.8. Statistical Risk Model Calibration – Risk decile plots or calibration curves:

2b3.9. Results of Risk Stratification Analysis:

2b3.10. 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*)

2b3.11. 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*)

2b4. IDENTIFICATION OF STATISTICALLY SIGNIFICANT & MEANINGFUL DIFFERENCES IN PERFORMANCE

2b4.1. 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)*

Descriptive statistics were constructed for each facility in the PCDB. The statistics included the mean, standard deviation, standard error of the means, median, range of rates, and the interquartile range of rates across the facilities. Meaningful differences were viewed as those facilities with a rate more than two standard deviations above or below the mean of the facilities in the PCDB.

2b4.2. 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 descriptive statistics are listed below:

Number of facilities: 87
Range of rates (performance): 0.0% - 13.9%
Mean: 4.8%
Standard Deviation: 3.1%
Standard Error: 0.32%
Median: 4.3%
IQR: 4.4%

2b4.3. 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?*)

Three facilities had a performance rate greater than two standard deviations above the mean for the dataset and we would interpret that they performed significantly worse than average.

2b5. 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 social risk 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 social risk 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.***

2b5.1. 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*)

2b5.2. 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*)

2b5.3. 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*)

2b6. MISSING DATA ANALYSIS AND MINIMIZING BIAS

2b6.1. 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*)

Not applicable. The PCDB dataset includes all encounters used in the measure numerator and denominator and validation is performed at the facility-level for each data submission.

2b6.2. 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 applicable

2b6.3. 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*)

Not applicable

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*) Update this field for **maintenance of endorsement**.

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. For **maintenance of endorsement**, if this measure is not an eMeasure (eCQM), please describe any efforts to develop an eMeasure (eCQM).

NA

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. Please also complete and attach the NQF Feasibility Score Card.

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. Required for maintenance of endorsement. Describe difficulties (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 instrument-based, consider implications for both individuals providing 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)*.

Measure is calculated using MS-DRG and ICD-10 code criteria

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.

Specific Plan for Use	Current Use (for current use provide URL)
Quality Improvement (Internal to the specific organization)	Public Reporting Leapfrog https://ratings.leapfroggroup.org/measure/hospital/episiotomies

4a1.1 For each CURRENT use, checked above (update for maintenance of endorsement), provide:

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

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.

4a1.2. 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?*)

4a1.3. 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.*)

4a2.1.1. Describe how performance results, data, and assistance with interpretation have been provided to those being measured or other users during development or implementation.

How many and which types of measured entities and/or others were included? If only a sample of measured entities were included, describe the full population and how the sample was selected.

A rolling four quarter rate is provided to all member and contracted hospitals (~ 150) every quarter. The data is calculated on 100% of eligible cases and the hospital´s rate is compared to a peer subgroup rate and the NPIC Data Base rate. The data are displayed in tabular and graphical form and hospitals may request case lists of those cases in the numerator for record auditing. Clinical experts are available should a hospital want guidance on how to reduce their rate.

4a2.1.2. Describe the process(es) involved, including when/how often results were provided, what data were provided, what educational/explanatory efforts were made, etc.

See above. Quarterly webinars provide member and contracted hospitals a chance to review all their data and request input from other hospitals either struggling or successful in moving their rate.

4a2.2.1. Summarize the feedback on measure performance and implementation from the measured entities and others described in 4d.1.

Describe how feedback was obtained.

Feedback indicates the rate is fairly and accurately calculated and reflective of their true performance. The tabular and graphical display are helpful to benchmark their performance against a peer subgroup. Feedback is obtained via webinar or individual contact with the hospital.

4a2.2.2. Summarize the feedback obtained from those being measured.

See above.

4a2.2.3. Summarize the feedback obtained from other users

Valid and helpful measure.

4a2.3. Describe how the feedback described in 4a2.2.1 has been considered when developing or revising the measure specifications or implementation, including whether the measure was modified and why or why not.

The measure has not changed except for coding and DRG updates indicating the measure is valuable perinatal metric to track.

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.

4b1. Refer to data provided in 1b but do not repeat here. Discuss any progress on improvement (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.)

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.

Looking at historical data from NPIC DB improvement has been shown year over year:

CY 2010: 11.5%

CY 2014: 7.2%

CY 2019: 4.7%

4b2. 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).

4b2.1. Please explain any unexpected findings (positive or negative) during implementation of this measure including unintended impacts on patients.

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 successfully 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 susceptibility 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 discrepancy, 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 episiotomies. The overall rate of coding error was on less 3%.

4b2.2. Please explain any unexpected benefits from implementation of 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.

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 of Related Measures

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 harmonized to the extent possible?

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.

No appendix **Attachment:**

Contact Information

Co.1 Measure Steward (Intellectual Property Owner): Christiana Care Health System

Co.2 Point of Contact: Matthew, Hoffman, mhoffman@christianacare.org, 302-301-3350-

Co.3 Measure Developer if different from Measure Steward: National Perinatal Information Center

Co.4 Point of Contact: Matthew, Hoffman, mhoffman@christianacare.org, 302-301-3350-

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? 11, 2020

Ad.6 Copyright statement:

Ad.7 Disclaimers:

Ad.8 Additional Information/Comments: