**National Quality Forum—Measure Testing (subcriteria 2a2, 2b2-2b7)**

**Measure Number** (*if previously endorsed*)**:** Click here to enter NQF number

**Measure Title**: Birthrisk Cesarean Birth Measure

**Date of Submission**: 2/12/2016

**Type of Measure:**

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

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

|  |
| --- |
| **Note: The information provided in this form is intended to aid the Steering Committee and other stakeholders in understanding to what degree the testing results for this measure meet NQF’s evaluation criteria for testing.**  **2a2.** **Reliability testing** [**10**](#Note10) demonstrates the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise. For **PRO-PMs and composite performance measures**, reliability should be demonstrated for the computed performance score.  **2b2.** **Validity testing** [**11**](#Note11) demonstrates that the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality. For **PRO-PMs and composite performance measures**, validity should be demonstrated for the computed performance score.    **2b3.** Exclusions are supported by the clinical evidence; otherwise, they are supported by evidence of sufficient frequency of occurrence so that results are distorted without the exclusion; [**12**](#Note12)  **AND**  If patient preference (e.g., informed decisionmaking) is a basis for exclusion, there must be evidence that the exclusion impacts performance on the measure; in such cases, the measure must be specified so that the information about patient preference and the effect on the measure is transparent (e.g., numerator category computed separately, denominator exclusion category computed separately). [**13**](#Note13)  **2b4.** **For outcome measures and other measures when indicated** (e.g., resource use):   * **an evidence-based risk-adjustment strategy** (e.g., risk models, risk stratification) is specified; is based on patient factors (including clinical and sociodemographic factors) that influence the measured outcome and are present at start of care; [**14**](#Note14)**,**[**15**](#Note15) and has demonstrated adequate discrimination and calibration   **OR**   * rationale/data support no risk adjustment/ stratification.   **2b5.** Data analysis of computed measure scores demonstrates that methods for scoring and analysis of the specified measure allow for **identification of statistically significant and practically/clinically meaningful** [**16**](#Note16) **differences in performance**;  **OR**  there is evidence of overall less-than-optimal performance.  **2b6.** **If multiple data sources/methods are specified, there is demonstration they produce comparable results**.  **2b7.** For **eMeasures, composites, and PRO-PMs** (or other measures susceptible to missing data),analyses identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias.  **Notes**  **10.** Reliability testing applies to both the data elements and computed measure score. Examples of reliability testing for data elements include, but are not limited to: inter-rater/abstractor or intra-rater/abstractor studies; internal consistency for multi-item scales; test-retest for survey items. Reliability testing of the measure score addresses precision of measurement (e.g., signal-to-noise).  **11.** Validity testing applies to both the data elements and computed measure score. Validity testing of data elements typically analyzes agreement with another authoritative source of the same information. Examples of validity testing of the measure score include, but are not limited to: testing hypotheses that the measures scores indicate quality of care, e.g., measure scores are different for groups known to have differences in quality assessed by another valid quality measure or method; correlation of measure scores with another valid indicator of quality for the specific topic; or relationship to conceptually related measures (e.g., scores on process measures to scores on outcome measures). Face validity of the measure score as a quality indicator may be adequate if accomplished through a systematic and transparent process, by identified experts, and explicitly addresses whether performance scores resulting from the measure as specified can be used to distinguish good from poor quality.  **12.** Examples of evidence that an exclusion distorts measure results include, but are not limited to: frequency of occurrence, variability of exclusions across providers, and sensitivity analyses with and without the exclusion.  **13.** Patient preference is not a clinical exception to eligibility and can be influenced by provider interventions.  **14.** Risk factors that influence outcomes should not be specified as exclusions  **15.** With large enough sample sizes, small differences that are statistically significant may or may not be practically or clinically meaningful. The substantive question may be, for example, whether a statistically significant difference of one percentage point in the percentage of patients who received smoking cessation counseling (e.g., 74 percent v. 75 percent) is clinically meaningful; or whether a statistically significant difference of $25 in cost for an episode of care (e.g., $5,000 v. $5,025) is practically meaningful. Measures with overall less-than-optimal performance may not demonstrate much variability across providers. |

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

Data used to create and validate this measure were obtained by application from the New York State Department of Health, Bureau of Biometrics and Health Statistics which contained non-identifiable information from 505,696 births certificate records from births that occurred in New York State exclusive of New York City between 2004 and 2007. Records were excluded if the birth occurred before 36 weeks 4 days or after 42 weeks 3 days of gestation (92,089), the patient had a history of a prior cesarean birth (52,833), was a multiple gestation (20,325), presentation was non-vertex (18,568) or if the patient did not attempt labor (17,797). There were 304,084 birth records that met the criteria for inclusion of which 19,928 (6.6%) were missing or had an unusual required data element resulting in 284,156 birth records from 2,915 different obstetrical care providers working out of 109 hospitals in the initial sample.

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

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

|  |  |
| --- | --- |
| **Measure Specified to Use Data From:**  **(*must be consistent with data sources entered in S.23*)** | **Measure Tested with Data From:** |
| abstracted from paper record | abstracted from paper record |
| administrative claims | administrative claims |
| clinical database/registry | clinical database/registry |
| abstracted from electronic health record | abstracted from electronic health record |
| eMeasure (HQMF) implemented in EHRs | eMeasure (HQMF) implemented in EHRs |
| other: Birth Certificate Records | other: Birth Certificate Records |

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

**1.3. What are the dates of the data used in testing**? 2004 - 2007

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

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

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

There were 2,915 obstetrical care providers working out of 109 facilities in the data obtained from New York State. This data included all obstetrical care providers and facilities in New York State exclusive of those within New York City.

**1.6. How many and which patients were included in the testing and analysis (by level of analysis and data source)**? (*identify the number and descriptive characteristics of patients included in the analysis (e.g., age, sex, race, diagnosis); if a sample was used, describe how patients were selected for inclusion in the sample*)

Of the 505,696 birth records from 2004-2007 that were obtained, all of the patients who attempted labor without a history of a prior cesarean birth and gave birth to a single baby in vertex presentation between 36 weeks 4 days and 42 weeks 3 days of gestation were included in the testing and analysis There were 304,084 birth records that met the criteria for inclusion of which 19,928 (6.6%) were missing or had an unusual required data element resulting in 284,156 birth records from 2,915 different obstetrical care providers working out of 109 hospitals in the initial sample. The number and characteristics of the data file are provided in the following three tables.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Year | 2004 | 2005 | 2006 | 2007 | Total |
| **Total records** | **127,489** | **125,297** | **126,446** | **126,464** | **505,696** |
|  |  |  |  |  |  |
| Non term gestations | 23,296 | 21,871 | 23,609 | 23,313 | 92,089 |
| Prior cesarean delivery | 12,627 | 12,604 | 12,625 | 14,977 | 52,833 |
| Multiple gestations | 5,204 | 4,813 | 5,028 | 5,280 | 20,325 |
| Non vertex presentations | 5,366 | 4,648 | 4,613 | 3,941 | 18,568 |
| Did not labor | 3,136 | 4,753 | 5,641 | 4,267 | 17,797 |
| Unknown delivery type | 394 | 97 | 110 | 31 | 632 |
| Bad facility code | 551 | 569 | 583 | 692 | 2,395 |
| Missing or unusual information | 4,138 | 3,935 | 4,882 | 3,946 | 16,901 |
|  |  |  |  |  |  |
| **Eligible for analysis** | **72,777** | **72,007** | **69,355** | **70,017** | **284,156** |
|  |  |  |  |  |  |
| Nulliparous records | 33,158 | 32,967 | 32,118 | 33,113 | 131,356 |
| Spontaneous nulliparous records | 24,508 | 23,464 | 22,795 | 23,146 | 93,913 |
| Induced nulliparous records | 8,650 | 9,503 | 9,323 | 9,967 | 37,443 |
|  |  |  |  |  |  |
| Multiparous records | 39,619 | 39,040 | 37,237 | 36,904 | 152,800 |
| Spontaneous multiparous records | 30,995 | 29,446 | 27,794 | 27,619 | 115,854 |
| Induced multiparous records | 8,624 | 9,594 | 9,443 | 9,285 | 36,946 |
|  |  |  |  |  |  |
| **Number of Hospitals** | **109** | **104** | **101** | **101** | **109** |
| **Number of Providers** | **2,107** | **2,031** | **1,896** | **1,881** | **2,915** |

|  |  |  |
| --- | --- | --- |
| **Nulliparous (N=131,356)** |  |  |
|  | **n (Spont./Induced)** | **Total (%)** |
| **Height (inches)** |  |  |
| <60 | 4,052 (2,983/1,069) | 3.08 |
| 60-62 | 31,889 (23,180/8,709) | 24.28 |
| 63-65 | 54,381 (39,026/15,355) | 41.40 |
| 66-68 | 33,525 (23,555/9,970) | 25.52 |
| >68 | 7,509 (5,169/2340) | 5.72 |
|  |  |  |
| **Fetal Weight (grams)** |  |  |
| <2500 | 3,302 (2,104/1,198) | 2.51 |
| 2500-2900 | 14,922 (10,890/4,032) | 11.36 |
| 2901-3300 | 38,981 (29,345/9,636) | 29.68 |
| 3301-3700 | 43,698 (31,458/12,240) | 33.27 |
| 3701-4100 | 22,955 (15,421/7,534) | 17.48 |
| 4101-4500 | 6,314 (3,973/2,341) | 4.81 |
| >4500 | 1,184 (722/462) | 0.90 |
|  |  |  |
| **Prepregnancy BMI (kg/m2)** |  |  |
| <20 | 12,187 (9,540/2,647) | 9.28 |
| 20-26 | 78,478 (58,367/20,111) | 59.74 |
| 27-33 | 29,033 (19,336/9,697) | 22.10 |
| 34-40 | 8,609 (5,051/3,558) | 6.55 |
| >40 | 3,049 (1,619/1,430) | 2.32 |
|  |  |  |
| **Age (years)** |  |  |
| <20 | 18,951 (14,111/4,840) | 14.43 |
| 20-26 | 47,541 (34,252/13,289) | 36.19 |
| 27-33 | 47,735 (33,917/13,818) | 36.34 |
| 34-40 | 15,902 (10,864/5,038) | 12.11 |
| >40 | 1,227 (769/458) | 0.93 |
|  |  |  |
| **Gestational Age (weeks)** |  |  |
| 37 | 5,207 (3,791/1,416) | 3.96 |
| 38 | 15,937 (11,829/4,108) | 12.13 |
| 39 | 30,198 (22,737/7,461) | 22.99 |
| 40 | 39,505 (29,602/9,903) | 30.07 |
| 41 | 30,206 (19,629/10,577) | 23.00 |
| 42 | 10,303 (6,325/3,978) | 7.84 |
|  |  |  |
| **Weight Gain (lbs.)** |  |  |
| <11 | 5,695 (3,965/1,730) | 4.34 |
| 11-29 | 42,489 (31,202/11,287) | 32.35 |
| 30-48 | 64,060 (46,223/17,837) | 48.77 |
| >48 | 19,112 (12,523/6,589) | 14.55 |

|  |  |  |
| --- | --- | --- |
| **Multiparous (N=152,800)** |  |  |
|  | **n (Spont./Induced)** | **Total (%)** |
| **Height (inches)** |  |  |
| <60 | 4,522 (3718/804) | 2.96 |
| 60-62 | 35,672 (28,120/7552) | 23.35 |
| 63-65 | 62,335 (47,227/15,108) | 40.80 |
| 66-68 | 40,750 (30,047/10,703) | 26.67 |
| >68 | 9,521 (6,742/2,779) | 6.23 |
|  |  |  |
| **Fetal Weight (grams)** |  |  |
| <2500 | 2,556 (1,792/764) | 1.67 |
| 2500-2900 | 13,041 (9,978/3,063) | 8.53 |
| 2901-3300 | 39,845 (31,180/8,665) | 26.08 |
| 3301-3700 | 52,356 (40,020/12,336) | 34.26 |
| 3701-4100 | 32,225 (23,864/8,361) | 21.09 |
| 4101-4500 | 10,509 (7,398/3,111) | 6.88 |
| >4500 | 2,268 (1,622/646) | 1.48 |
|  |  |  |
| **Prepregnancy BMI (kg/m2)** |  |  |
| <20 | 11,752 (9,532/2,220) | 7.69 |
| 20-26 | 85,516 (66,818/18,698) | 55.97 |
| 27-33 | 39,661 (29,097/10,564) | 25.96 |
| 34-40 | 11,919 (8,006/3,913) | 7.80 |
| >40 | 3,952 (2,401/1,551) | 2.59 |
|  |  |  |
| **Age (years)** |  |  |
| <20 | 2,748 (2,244/504) | 1.80 |
| 20-26 | 40,564 (31,556/9,008) | 26.55 |
| 27-33 | 64,467 (48,618/15,849) | 42.19 |
| 34-40 | 41,198 (30,719/10,479) | 26.96 |
| >40 | 3,823 (2,717/1,106) | 2.50 |
|  |  |  |
| **Gestational Age (weeks)** |  |  |
| 37 | 6,689 (5,254/1,435) | 4.38 |
| 38 | 22,503 (17,285/5,218) | 14.73 |
| 39 | 42,574 (32,413/10,161) | 27.86 |
| 40 | 45,455 (35,466/9,989) | 29.75 |
| 41 | 26,998 (19,347/7,651) | 17.67 |
| 42 | 8,581 (6,089/2,492) | 5.62 |
|  |  |  |
| **Weight Gain (lbs.)** |  |  |
| <11 | 11,530 (8,587/2,943) | 7.55 |
| 11-29 | 64,596 (49,602/14,994) | 42.27 |
| 30-48 | 64,709 (49,230/15,479) | 42.35 |
| >48 | 11,965 (8,435/3,530) | 7.83 |

**1.7. If there are differences in the data or sample used for different aspects of testing (e.g., reliability, validity, exclusions, risk adjustment), identify how the data or sample are different for each aspect of testing reported below**.

No differences

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

Sociodemographic variables were not collected or analyzed.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**2a2. RELIABILITY TESTING**

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

**2a2.1. What level of reliability testing was conducted**? (*may be one or both levels*)  
 **Critical data elements used in the measure** (*e.g., inter-abstractor reliability; data element reliability must address ALL critical data elements*)  
 **Performance measure score** (e.g., *signal-to-noise analysis*)  
  
**2a2.2. For each level checked above, describe the method of reliability testing and what it tests** (*describe the steps―do not just name a method; what type of error does it test; what statistical analysis was used*)

In 2006 Northam and Knapp published a review of twenty four primary research studies of U.S. birth certificates that involved validity or reliability assessment. They reported that: “Reliability is admittedly difficult to evaluate for birth certificate data. The dichotomous data are not amenable to traditional internal consistency evaluation because birth certificates provide data about individual variables rather than dimensions of a single construct, and they use a single item for each variable. Because birth certificate data reflect a cross-sectional assessment of each birth and are concerned with variables such as birthweight, obstetric care, and the presence of infant abnormalities that are reported at one time, they are also not amenable to an investigation of temporal stability. Thus, both internal consistency and stability assessments are not feasible methods of evaluating the reliability of birth certificate data.”

See section 2b2 for validity testing of data elements

A signal to noise analysis for the hospital performance measure score was conducted to provide reliability testing for the performance measure score. MS group is the signal (2.532). MS error is the noise (27.910). This will test if the between-group variance (signal) is comparable to the within-group variance (noise). Signal-to-noise ratio was calculated as follows: 2.532/27.910 = 0.09072

A signal to noise analysis for the provider performance measure score was conducted to provide reliability testing for the performance measure score. MS group is the signal (777.361). MS error is the noise (709.237). This will test if the between-group variance (signal) is comparable to the within-group variance (noise). Signal-to-noise ratio was calculated as follows: 777.361/709.237= 1.096053

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

Hospital Performance Measure Score:

*ANOVA*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Sum of Squares | df | Mean Square | F | Sig. |
| Between Groups | 5.063 | 2 | 2.532 | .091 | .913 |
| Within Groups | 8456.786 | 303 | 27.910 |  |  |
| Total | 8461.849 | 305 |  |  |  |

Provider Performance Measure Score:

*ANOVA*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Sum of Squares | df | Mean Square | F | Sig. |
| Between Groups | 1554.722 | 2 | 777.361 | 1.096 | .334 |
| Within Groups | 4118540.381 | 5807 | 709.237 |  |  |
| Total | 4120095.102 | 5809 |  |  |  |

**2a2.4 What is your interpretation of the results in terms of demonstrating reliability**? (i*.e., what do the results mean and what are the norms for the test conducted?*)

A ratio higher than 1:1 (greater than 0 dB) indicates more signal than noise. Thus, the between-group variance (signal) is comparable to the within-group variance. The Cronbach’s alpha was .87 that was greater than .70 (Nyunnally, 1978) indicating high reliability.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**2b2. VALIDITY TESTING**

**2b2.1. What level of validity testing was conducted**? (*may be one or both levels*)  
 **Critical data elements** (*data element validity must address ALL critical data elements*)

**Performance measure score**

**Empirical validity testing** **Systematic assessment of face validity of performance measure score as an indicator** of quality or resource use (*i.e., is an accurate reflection of performance on quality or resource use and can distinguish good from poor performance*)

**2b2.2. For each level of testing checked above, describe the method of validity testing and what it tests** (*describe the steps―do not just name a method; what was tested, e.g., accuracy of data elements compared to authoritative source, relationship to another measure as expected; what statistical analysis was used)*

Northam and Knapp reported that “The 1st U.S. Standard Certificate of Live Birth was developed in 1900 ( Gould, 1999 ), and birth certificates have included data on maternal and infant variables for all live births in the United States and its territories since 1939 ( Lilienfeld, Parkhurst, Patton, & Schlesinger, 1951 ).” Despite the long history regarding the use of the birth certificate, validity testing of the data elements provided by birth certificate data remains elusive with most studies regarding the validity of birth certificate data elements being conducted prior the 2003 revision of the U.S. birth certificate. Recently Martin et al. published a National Vital Statistics Report in 2013 comparing the data elements in 600 birth certificate records in one state and 495 birth certificate records in another to the information recorded in the medical record. Each data element in the medical record was compared to the birth certificate data element and the percentage of exact matches was recorded.

Performance measure score validity testing was conducted using construct validity for the hospital performance measure score and criterion validity for the provider performance measure score. The type of criterion validity that was used was predictive validity.

**2b2.3. What were the statistical results from validity testing**? (*e.g., correlation; t-test*)

Validity of the data elements was examined in the National Vital Statistics Report from 2013 which compared many but not all of the birth certificate data elements to the results found in the medical record and reported the findings based on the percentage of exact agreement as follows:

|  |  |  |
| --- | --- | --- |
|  | Percent Agreement | |
| U. S. Birth Certificate Item | State A | State B |
|  |  |  |
| Number of previous live births now living | 96.0% | 96.1% |
| Number of previous live births now dead | 99.2% | 97.9% |
| Number of previous cesarean deliveries | 95.3% | 92.5% |
| Obstetric estimate of gestation at delivery (exact weeks) | 91.6% | 67.4% |
| Obstetric estimate of gestation (within 2 weeks) | 99.7% | 98.1% |
| Birthweight (exact grams) | 90.0% | 90.7% |
| Birthweight within 500 grams | 99.7% | 99.4% |
| Induction of labor | 89.7% | 83.0% |
| Cephalic Presentation | 92.0% | 93.2% |
| Breech Presentation | 98.2% | 98.0% |
| Cesarean Birth | 98.8% | 97.2% |
| Trial of labor attempted | 89.0% | 84.6% |

An older report from Querec (1980) comparing birth certificate data with maternal survey data revealed that excellent agreement (90% or better) was noted for maternal age and plurality.

The performance measure score validity for the hospital performance measure score revealed that the Cronbach’s alpha was .87. The performance measure score predictive validity for the provider performance measure score revealed a regression line with an equation of y = 1.017x + 1.4964 and an R-squared value of 0.9802.

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

The data elements that have been studied were demonstrated to have high validity when compared to the medical record with the exception of Induction of Labor and Trial of labor which were deemed to have substantial validity. There was a discrepancy in Obstetric estimate of gestation at delivery (exact weeks) between the two states with state B only having moderate validity.

Unfortunately not all of the data elements that are relevant to this measure have been tested for validity. The data elements of Mother’s Height, Mother’s Prepregnancy Weight and Mother’s Weight at Delivery were not addressed any in of the studies. Further testing will be required to determine if validity of these data elements is similar to the validity of the tested data elements.

The Cronbach’s alpha was 0.87, which was greater than .70 (Nyunnally, 1978). High reliability implies high construct validity. The alpha coefficient of the performance measure score was greater than .70 (Nyunnally, 1978). Thus, the hospital performance measure score has high construct validity.

The performance measure score predictive validity testing reveals an excellent correlation between the actual cesarean birth rate to the predicted cesarean birth rate using the provider performance measure score.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**2b3. EXCLUSIONS ANALYSIS**

**NA**  **no exclusions — *skip to section*** [***2b4***](#section2b4)

**2b3.1. Describe the method of testing exclusions and what it tests** (*describe the steps―do not just name a method; what was tested, e.g., whether exclusions affect overall performance scores; what statistical analysis was used*)

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

**2b3.3. What is your interpretation of the results in terms of demonstrating that exclusions are needed to prevent unfair distortion of performance results?** (*i.e., the value outweighs the burden of increased data collection and analysis.*  *Note:* ***If patient preference is an exclusion****, the measure must be specified so that the effect on the performance score is transparent, e.g., scores with and without exclusion*)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**2b4. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES**  
***If not an intermediate or health outcome, or PRO-PM, or resource use measure, skip to section*** [***2b5***](#section2b5)***.***

**2b4.1. What method of controlling for differences in case mix is used?**

**No risk adjustment or stratification**

**Statistical risk model with** Click here to enter number of factors **risk factors**

**Stratification by** Click here to enter number of categories **risk categories**

**Other,** Cohort comparison

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

N/A

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

The factors selected for the risk adjustment were: parity (nulliparous or multiparous), newborn weight, prepregnancy maternal body mass index, maternal age, maternal height, gestational age, maternal weight gain, and the type of onset of labor (spontaneous or induced). These factors were selected because they are objective, are typically measured in every pregnancy and have been previously proven to significantly affect the risk that labor will result in a cesarean birth.

The selected factors were confirmed to significantly affect the risk that labor will result in a cesarean birth in the following manner. Parity and onset of labor were analyzed by z-test. Newborn weight, prepregnancy maternal body mass index, maternal age, maternal height, gestational age and maternal weight gain were analyzed using multivariate logistic regression analysis.

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

Analysis of the 284,156 birth records revealed that 32,540 (24.77%) of the 131,356 nulliparous women had a cesarean birth and 7,679 (5.03%) of the 152,800 multiparous women had a cesarean birth. The cesarean birth rate in nulliparous women was significantly higher than in multiparous women (p<=0.001). Analysis of the 131,356 nulliparous birth records revealed that 13,558 (36.21%) of the 37,443 women whose labor was induced had a cesarean birth and 18,982 (20.12%) of the 93,913 women whose labor was spontaneous had a cesarean birth. The cesarean birth rate for nulliparous women whose labor was induced was significantly higher than in nulliparous women whose labor was spontaneous (p<=0.001). Analysis of the 152,800 multiparous birth records revealed that 2,749 (7.44%) of the 36,946 women whose labor was induced had a cesarean birth and 4,930 (4.26%) of the 115,854 women whose labor was spontaneous had a cesarean birth. The cesarean birth rate for multiparous women whose labor was induced was significantly higher than in multiparous women whose labor was spontaneous (p<=0.001).

Both newborn weight and maternal weight gain did not exhibit a linear progression throughout the reported ranges. Increasing newborn weight when over 2900 grams, increasing prepregnancy body mass index, increasing maternal age, decreasing maternal height, decreasing newborn weight when under 2900 grams, longer gestations, increasing maternal weight gain when over 25 pounds and decreasing maternal weight gain when under 25 pounds significantly affected the risk for cesarean birth for both nulliparous and multiparous women (p<=0.001). The effect of each physical characteristic on the rate of cesarean birth for the 284,156 women is illustrated in figures 1 through 6 and in the two tables that follow.

### 

### Figure 1 – Effect of newborn weight Figure 2 – Effect of prepregnancy body mass index

**

**Figure 3 – Effect of maternal age** **Figure 4 – Effect of maternal height**

**Figure 5 – Effect of gestational age Figure 6 – Effect of maternal weight gain**

**Adjusted odds ratio progression of cesarean birth for nulliparous women:**

|  |  |  |  |
| --- | --- | --- | --- |
| Physical characteristic | Step Size | Odds Ratio | 95% Confidence Interval |
| Newborn weight when > 2900 g | Every 200 g larger | 1.331 | 1.310-1.352 |
| Prepregnancy maternal body mass index | Every 3 kg/m2 increase | 1.210 | 1.204-1.219 |
| Maternal age | Every 3 years older | 1.195 | 1.186-1.201 |
| Maternal height | Every inch shorter | 1.124 | 1.119-1.128 |
| Newborn weight when <= 2900 g | Every 200 g smaller | 1.122 | 1.107-1.134 |
| Gestational age | Every week longer | 1.100 | 1.088-1.113 |
| Maternal weight gain when > 25 lbs | Every 5 lbs more | 1.078 | 1.070-1.085 |
| Maternal weight gain when <= 25 lbs | Every 5 lbs less | 1.020 | 1.005-1.035 |

**Adjusted odds ratio progression of cesarean birth for multiparous women:**

|  |  |  |  |
| --- | --- | --- | --- |
| Physical characteristic | Step Size | Odds Ratio | 95% Confidence Interval |
| Prepregnancy body mass index | Every 3 kg/m2 increase | 1.201 | 1.189-1.210 |
| Newborn weight when <= 2900 g | Every 200 g smaller | 1.155 | 1.140-1.165 |
| Maternal age | Every 3 years older | 1.153 | 1.141-1.165 |
| Newborn weight when > 2900 g | Every 200 g larger | 1.149 | 1.127-1.171 |
| Maternal height | Every inch shorter | 1.091 | 1.083-1.099 |
| Maternal weight gain when > 25 lbs | Every 5 lbs more | 1.085 | 1.075-1.095 |
| Gestational age | Every week longer | 1.035 | 1.014-1.055 |
| Maternal weight gain when <= 25 lbs | Every 5 lbs less | 1.030 | 1.005-1.055 |

**2b4.4b. Describe the analyses and interpretation resulting in the decision to select SDS factors (e.g. prevalence of the factor across measured entities, empirical association with the outcome, contribution of unique variation in the outcome, assessment of between-unit effects and within-unit effects)**

**SDS factors were not selected.**

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

Validation of the accuracy of the risk adjustment is achieved by using the risk adjustment to predict the outcomes for women whose births were not used to create the risk adjustment. The inherent risk previously assigned to each of 211,379 test records (those from 2005 to 2007) along with the obstetrical care provider’s risk adjustment at the time of that birth is used to calculate the predicted outcome for that test record. Predictions were limited to a range from 1% to 99%. However, if the obstetrical care provider did not yet have ten birth records in the database then that test record was not assigned a predicted risk.

This resulted in 182,757 birth records for which a predicted risk of cesarean birth was calculated. The 182,757 records were grouped according to their predicted risk from 1% to 99%. Ten of these groups had less than ten test records in the group and were not used in the comparison. The predicted risk groups that were omitted from the comparison were 88%, 90%, 91% and 93% to 99% inclusive. This resulted in 182,722 records for comparison. The actual cesarean birth rate for each group was compared to the predicted cesarean birth rate for each group.

*Provide the statistical results from testing the approach to controlling for differences in patient characteristics (case mix) below*.  
***If stratified, skip to*** [***2b4.9***](#question2b49)

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



The comparison of the actual cesarean birth rate to the predicted cesarean birth rate resulted in a regression line with an equation of y = 1.017x + 1.4964 and an R-squared of 0.9802.

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

The Hosmer-Lemeshow test statistic is 348.34 (df=8; p-value=0.00).

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

**2b4.9. Results of Risk Stratification Analysis**:

**2b4.10. What is your interpretation of the results in terms of demonstrating adequacy of controlling for differences in patient characteristics (case mix)?** (i*.e., what do the results mean and what are the norms for the test conducted*)

Validation of the risk adjustment was obtained by predicting the outcome for births that were not included in assigning each obstetrical care provider’s risk adjustment. The large number of predictions in the validation allowed for a percentage point by percentage point analysis which revealed an R-squared value of 0.9802. This high R-squared value reveals the accuracy of both the inherent risk assigned to each birth and the risk adjustment assigned to the obstetrical care provider.

In the validation, the obstetrical care provider’s risk adjustment from their previous 100 births is used to assign the predicted risk to each subsequent birth. Therefore, if the cesarean birth rate is changing over time the predicted outcomes may underestimate or overestimate the actual outcomes. The cesarean birth rate increased from 13.45% in 2004 to 14.12%, 14.19% and 14.89% in the following three years. As expected with an increasing cesarean birth rate the predictions are slightly underestimated. The large sample size and the underestimation due to the increasing cesarean birth rate may explain the failure of the Hosmer-Lemeshow test. A changing cesarean birth rate only affects the validation of the risk adjustment and not the ability to compare cesarean birth measures between entities for a specific period of time.

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

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

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

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

The Birthrisk Cesarean Birth Measure provides the ability to assess differences in performance at several levels. The obstetrical care provider can compare themselves to the other providers in their facility, county, state or across the nation. A facility can compare themselves to the other facilities in their county, state or across the nation. Healthcare administrators can compare results at the level of the obstetrical care provider, facility, county, state or across the nation. A simple z-test is used to create a 95% confidence interval to determine if the selected cesarean birth measure is statistically significantly different than that of another entity or that of the average in the facility, county, state or nation.

**2b5.2. What were the statistical results from testing the ability to identify statistically significant and/or clinically/practically meaningful differences in performance measure scores across measured entities?** (e.g., *number and percentage of entities with scores that were statistically significantly different from mean or some benchmark, different from expected; how was meaningful difference defined*)

There were 211,379 birth records (2005-2007) from 2,542 different obstetrical care providers working out of 105 different facilities in New York State that were used to assess performance in the initial sample.

Comparing the 2,542 obstetrical care providers to the average cesarean birth measure for the state, there were 1,802 (70.9%) obstetrical care providers that had a cesarean birth measure that was not significantly different than the average and they represented 55% of the total births. There were 398 (15.6%) obstetrical care providers that had a cesarean birth measure that was significantly better than the average and they represented 26% of the total births. There were 342 (13.5%) obstetrical care providers that had a cesarean birth measure that was significantly worse than the average and they represented 19% of the total births.

Comparing the 105 facilities to the average cesarean birth measure for the state, there were 42 (40%) facilities that had a cesarean birth measure that was not significantly different than the average and they represented 28.6% of the total births. There were 30 (29%) facilities that had a cesarean birth measure that was significantly better than the average and they represented 38.0% of the total births. There were 33 (31%) facilities that had a cesarean birth measure that was significantly worse than the average and they represented 33.4% of the total births.

Comparing the 86 obstetrical care providers who work in the New York State facility with the most births (9,364) to the average cesarean birth measure for the state, there were 72 (84%) obstetrical care providers that had a cesarean birth measure that was not significantly different than the average and they represented 71% of the total births. There were 5 (9%) obstetrical care providers that had a cesarean birth measure that was significantly better than the average and they represented 9% of the total births. There were 9 (10%) obstetrical care providers that had a cesarean birth measure that was significantly worse than the average and they represented 18% of the total births. This facility was one of the 33 that had a cesarean birth measure that was worse than the average for the state.

**2b5.3. What is your interpretation of the results in terms of demonstrating the ability to identify statistically significant and/or clinically/practically meaningful differences in performance across measured entities?** (i*.e., what do the results mean in terms of statistical and meaningful differences?*)

The decision to perform a cesarean birth is made by obstetrical care providers and not by facilities. Therefore, the level of analysis required to improve outcome must be down to the level of the obstetrical care provider. For example, in the analysis of the facility with the most births it is the 9 obstetrical care providers in that facility who are performing 18% of the births that results in that facility being one of the 33 facilities that was found to have a cesarean birth measure that was significantly worse than the average for the state.

The Birthrisk Cesarean Birth Measure uses a simple z-test to create statistically significant comparisons between entities. The ability to report an obstetrical care provider’s cesarean birth measure as a comparison to the average in their facility, county, state or in the nation with a *p* value gives obstetrical care providers results in a manner with which they are familiar. Obstetrical care providers and facilities would have fewer objections to public reporting of their cesarean birth measure if the reporting is limited to providing results as better than average, average or worse than average based on a 95% confidence interval.

Reporting results based on a 95% confidence interval would also allow for an acceptable plan for value based incentive/disincentive payments. Paying providers or facilities more if they are significantly better than average and paying providers or facilities less if they are significantly worse than average will provide the correct incentive for the management of laboring patients because providers who perform more cesarean births than expected actually work less than those who perform less cesarean births than expected. This is because it is less work for the obstetrical care provider to end a labor by proceeding to a cesarean birth than it is to allow the labor to continue.

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

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

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

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

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

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

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

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**2b7. MISSING DATA ANALYSIS AND MINIMIZING BIAS**

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

Birth records in the initial sample were analyzed using IBM SPSS Statistical software version 23 and Little’s MCAR test was performed. Analysis included unconditional testing to find records that were missing any of the data elements or had data elements that were outside of the usual accepted range. After the unconditional tests of known requirements were performed, critical data elements were subjected to conditional tests to assure that the logical relationships exist and were appropriate. Listwise deletion of records with missing or unusual data elements was performed. The dataset of deleted records was compared to the remaining data sample. The distribution of deleted records by providers was analyzed. Bias is minimized by confirming that deleted records were similar to retained records. The ability to re-abstract data in future samples will result in less missing values which in turn will further reduce any potential bias.

**2b7.2. What is the overall frequency of missing data, the distribution of missing data across providers, and the results from testing related to missing data?** (*e.g.,**results of sensitivity analysis of the effect of various rules for missing data/nonresponse; if no empirical sensitivity analysis, identify the approaches for handling missing data that were considered and pros and cons of each*)

There were 224,142 records that met the criteria for inclusion in order to assess performance. Of these records there were 12,763 (5.69%) that were missing one or more of the critical data elements or had unusual data. Little’s MCAR test revealed that the missing values were not random. Analysis of the data records that were removed from the data sample revealed that 1,229 providers had at least one record deleted. For the 1,229 providers who had records deleted the average percentage of deleted records was 9.85%. Analysis of the means for the data elements revealed:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Parity -  Nulliparous | Induced | Baby Wt. | Prepreg Wt. | Postpreg Wt. | Age | Height | Weeks |
| Deleted Records | 43.79% | 20.01% | 3406 | 149.99 | 180.79 | 28.09 | 64.18 | 39.71 |
| Retained Records | 46.46% | 27.02% | 3421 | 150.48 | 182.07 | 28.31 | 64.24 | 39.68 |

The cesarean birth rate of the deleted records was 14.1% which was not statistically significantly different from the 14.4% cesarean birth rate for the 211,379 birth records used to assess performance.

Missing data was handled with Listwise deletion. Imputation methods were specifically not used as the performance measure is dependent on the unique combination of physical characteristics of the mother and the size of her baby. Improving data collection through re-abstraction will improve results.

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

The overall amount of deleted records is small and analysis reveals no significant difference in the physical characteristics of the deleted and retained records. The small difference in parity and induction of deleted records is unlikely to introduce bias because of the cohort comparison method that is used to create the risk adjustment for the performance measure. Since reducing the number of missing values provides the best method for reducing any potential bias, future data collections will benefit from re-abstraction, mandatory data element completion and verification of unusual data elements.