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

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

**Measure Title**: Perioperative Temperature Management

**Date of Submission**: 1/14/2015

**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). |

|  |
| --- |
| **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 that influence the measured outcome (but not factors related to disparities in care or the quality of care) 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.** Risk models should not obscure disparities in care for populations by including factors that are associated with differences/inequalities in care, such as race, socioeconomic status, or gender (e.g., poorer treatment outcomes of African American men with prostate cancer or inequalities in treatment for CVD risk factors between men and women). It is preferable to stratify measures by race and socioeconomic status rather than to adjust out the differences.  **16.** With large enough sample sizes, small differences that are statistically significant may or may not be practically or clinically meaningful. The substantive question may be, for example, whether a statistically significant difference of one percentage point in the percentage of patients who received smoking cessation counseling (e.g., 74 percent v. 75 percent) is clinically meaningful; or whether a statistically significant difference of $25 in cost for an episode of care (e.g., $5,000 v. $5,025) is practically meaningful. Measures with overall less-than-optimal performance may not demonstrate much variability across providers. |

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

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

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

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

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

The Anesthesia Quality Insitute National Anesthesia Clinical Outcome Registry (NACOR) Public Use File was used to assess available data on the submitted outcome Perioperative Temperature Management measure and the measure’s performance, as extracted according to the measure specifications, for 2010-2013.

For comparison on reliability, performance rates and other indicators of quality, The Anesthesia Quality Insitute National Anesthesia Clinical Outcome Registry (NACOR) Public Use File was used to assess data on the process Perioperative Tempaterature Management measure (NQF #0454/PQRS #193).

**1.3. What are the dates of the data used in testing**? 2010-2013

**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*)

For the outcome measure, cases meeting the measure criteria and where temperature was recorded have been included. For the process measure (NQF #0454/PQRS #193), cases for performance of the measure were identified by a provider completing a checkbox.

**Perioperative Temperature Management (Outcome Measure)**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **All cases in NACOR Public Use File (PUF) Q3 2014 where known providers are MD/DO or CRNAs[[1]](#footnote-1)** | | | | | |
|  | 2010 | 2011 | 2012 | 2013 | Total |
| Number of Practices | 2 | 3 | 2 | 1 | 4 |
| Number of Facilities | 2 | 3 | 2 | 1 | 4 |
| Number of Providers | 238 | 176 | 168 | 140 | 232 |
| Number of Cases | 1,628 | 388 | 3,884 | 4,690 | 10,590 |

**Perioperative Temperature Management (Process Measure – NQF #0454/PQRS #193)**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **All cases in NACOR Public Use File (PUF) Q3 2014 where known providers are MD/DO or CRNAs1** | | | | | |
|  | 2010 | 2011 | 2012 | 2013 | Total |
| Number of Practices | 18 | 40 | 57 | 99 | 102 |
| Number of Facilities | 80 | 160 | 287 | 492 | 559 |
| Number of Providers | 893 | 1,578 | 2,251 | 4,750 | 5,425 |
| Number of Cases | 32,690 | 89,548 | 155,590 | 247,951 | 525,779 |

**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*)

For the outcome Perioperative Temperature Management measure submitted, data was tested and analyzed on a provider and case basis where known providers were physicians or nurse anesthetists. Percentages may not add to 100% due to rounding.

| **Patient Age (Perioperative Temperature Management – Outcome)** | | | | |
| --- | --- | --- | --- | --- |
| **Age Group** | **2010 (%)** | **2011 (%)** | **2012 (%)** | **2013 (%)** |
| < 1 | 78 (4.79) | 10 (2.58) | 1 (0.03) | 3 (0.06) |
| 1 - 18 | 821 (50.43) | 155 (39.95) | 94 (2.42) | 133 (2.84) |
| 19 - 49 | 288 (17.69) | 97 (25.00) | 1,403 (36.12) | 1,690 (36.03) |
| 50 - 64 | 209 (12.84) | 53 (13.66) | 1,148 (29.56) | 1,448 (30.87) |
| 65 - 79 | 174 (10.69) | 56 (14.43) | 1,009 (25.98) | 1,187 (25.31) |
| 80+ | 58 (3.56) | 17 (4.38) | 229 (5.90) | 229 (4.8) |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Patient Sex (Perioperative Temperature Management – Outcome)** | | | | |
| Sex | **2010 (%)** | **2011 (%)** | **2012 (%)** | **2013 (%)** |
| Female | 766 (47.05) | 193 (49.74) | 2,113 (54.40) | 2,596 (55.35) |
| Male | 862 (52.95) | 195 (50.26) | 1,771 (45.60) | 2,094 (44.65) |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **ASA Physical Status (Perioperative Temperature Management – Outcome)** | | | | |
| ASA Physical Status | **2010 (%)** | **2011 (%)** | **2012 (%)** | **2013 (%)** |
| I - II | 6 (0.37) | 6 (1.55) | 1,761 (45.34) | 2,202 (46.95) |
| III | 2 (0.12) | 3 (0.77) | 1,702 (43.82) | 2,198 (46.87) |
| IV | 0 (0.00) | 0 (0.00) | 218 (5.61) | 241 (5.14) |
| V | 0 (0.00) | 0 (0.00) | 1 (0.03) | 1 (0.02) |
| Not Reported | 1,620 (99.51) | 379 (97.68) | 202 (5.20) | 48 (1.02) |

For comparison, ASA tested measure NQF #0454/PQRS #193 for the Medicare population.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Patient Age (Perioperative Temperature Management NQF #0454/PQRS #193– Process)** | | | | |
| Age Group | **2010 (%)** | **2011 (%)** | **2012 (%)** | **2013 (%)** |
| < 1 | 1 (<0.01) | 3 (<0.01) | 7 (<0.01) | 25 (0.01) |
| 1 - 18 | 3 (0.01) | 26 (0.03) | 39 (0.03) | 90 (0.04) |
| 19 - 49 | 2,054 (6.28) | 4,221 (4.71) | 7,403 (4.76) | 13,290 (5.36) |
| 50 - 64 | 4,144 (12.68) | 8,814 (9.84) | 15,762 (10.13) | 28,274 (11.40) |
| 65 - 79 | 18,329 (56.07) | 45,388 (50.69) | 83,970 (53.97) | 146,945 (59.26) |
| 80+ | 8,154 (24.94) | 19,474 (21.75) | 33,257 (21.38) | 55,909 (22.55) |
| Not Reported | 5 (0.02) | 11,622 (12.98) | 15,152 (9.74) | 3,418 (1.38) |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Patient Sex (Perioperative Temperature Management NQF #0454/PQRS #193 – Process)** | | | | |
| Sex | **2010 (%)** | **2011 (%)** | **2012 (%)** | **2013 (%)** |
| Female | 17,928 (54.84) | 42,887 (47.89) | 77,193 (49.61) | 134,079 (54.08) |
| Male | 14,761 (45.15) | 35,045 (39.14) | 63,237 (40.64) | 110,460 (44.55) |
| Not Reported | 1 (<0.01) | 11,616 (12.97) | 15,160 (9.74) | 3,412 (1.38) |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **ASA Physical Status (Perioperative Temperature Management NQF #0454/PQRS #193 – Process)** | | | | |
| ASA Physical Status | **2010 (%)** | **2011 (%)** | **2012 (%)** | **2013 (%)** |
| I - II | 14,951 (45.74) | 46,346 (51.76) | 73,859 (47.47) | 101,592 (40.97) |
| III | 13,058 (39.95) | 32,833 (36.72) | 62,400 (40.11) | 116,622 (47.03) |
| IV | 3,936 (12.04) | 7,191 (8.03) | 12,773 (8.21) | 25,860 (10.43) |
| V | 101 (0.31) | 157 (0.18) | 247 (0.16) | 465 (0.19) |
| Not Reported | 644 (1.97) | 2,971 (3.32) | 6,311 (4.06) | 3,412 (1.38) |

**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**.

There are no differences in the testing samples.

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

**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*)

Reliability was calculated according to the methods outlined in a technical report prepared by J.L. Adams titled “The Reliability of Provider Profiling: A Tutorial” (RAND Corporation, TR-653-NCQA, 2009). In this context, reliability represents the ability of a measure to confidently distinguish the performance of one physician from another. As discussed in the report: “Conceptually, it is the ratio of signal to noise. The signal in this case is the proportion of variability in measured performance that can be explained by real differences in performance. There are 3 main drivers of reliability; sample size, differences between physicians, and measurement error.”

According to this approach, reliability is estimated with a beta-binomial model. The beta-binomial model is appropriate for measuring the reliability of pass/fail measures such as those proposed.

**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*)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Perioperative Temperature Management (Outcome)[[2]](#footnote-2)** | | | |
|  | **2010** | **2011** | **2012** | **2013** |
|  |  |  |  |  |
| Number of Providers | 219 | 141 | 164 | 138 |
| Number of Cases | 1,504 | 364 | 3,806 | 4,576 |
| # provider-cases | 3,496 | 813 | 11,844 | 14,536 |
|  |  |  |  |  |
| **Performance** |  |  |  |  |
| Mean | 0.733 | 0.823 | 0.972 | 0.975 |
| Standard deviation | 0.189 | 0.219 | 0.036 | 0.037 |
| Interquartile range | 0.375 | 0.250 | 0.042 | 0.021 |
| Minimum | 0.143 | 0.000 | 0.750 | 0.750 |
| 10 %tile | 0.500 | 0.500 | 0.935 | 0.943 |
| 20 %tile | 0.571 | 0.667 | 0.948 | 0.959 |
| 30 %tile | 0.667 | 0.750 | 0.961 | 0.977 |
| 40 %tile | 0.706 | 0.857 | 0.972 | 0.980 |
| 50 %tile - Median | 0.765 | 1.000 | 0.982 | 0.986 |
| 60 %tile | 0.800 | 1.000 | 0.988 | 0.989 |
| 70 %tile | 0.857 | 1.000 | 1.000 | 0.992 |
| 80 %tile | 0.895 | 1.000 | 1.000 | 1.000 |
| 90 %tile | 1.000 | 1.000 | 1.000 | 1.000 |
| Maximum | 1.000 | 1.000 | 1.000 | 1.000 |
|  |  |  |  |  |
| **Reliability** |  |  |  |  |
| Mean of providers | 0.523 | 0.661 | 0.466 | 0.644 |
| Case-weighted | 0.672 | 0.653 | 0.413 | 0.696 |
| With exclusions: |  |  |  |  |
| Mean of providers | 0.527 | 0.611 | 0.424 | 0.531 |
| Case-weighted | 0.671 | 0.604 | 0.424 | 0.475 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Perioperative Temperature Management (Process): NQF #0454 / Measure #193[[3]](#footnote-3)** | | | |
|  | **2010** | **2011** | **2012** | **2013** |
|  |  |  |  |  |
| Number of Providers | 552 | 1,563 | 2,409 | 4,930 |
| Number of Cases | 32,692 | 89,549 | 155,593 | 247,956 |
| # provider-cases | 40,170 | 120,543 | 201,176 | 366,493 |
|  |  |  |  |  |
| **Performance** |  |  |  |  |
| Mean | 0.915 | 0.958 | 0.961 | 0.954 |
| Standard deviation | 0.201 | 0.138 | 0.092 | 0.115 |
| Interquartile range | 0.021 | 0.000 | 0.028 | 0.029 |
| Minimum | 0.000 | 0.000 | 0.000 | 0.000 |
| 10 %tile | 0.549 | 0.936 | 0.864 | 0.847 |
| 20 %tile | 0.957 | 0.995 | 0.949 | 0.946 |
| 30 %tile | 1.000 | 1.000 | 0.986 | 0.987 |
| 40 %tile | 1.000 | 1.000 | 1.000 | 1.000 |
| 50 %tile - Median | 1.000 | 1.000 | 1.000 | 1.000 |
| 60 %tile | 1.000 | 1.000 | 1.000 | 1.000 |
| 70 %tile | 1.000 | 1.000 | 1.000 | 1.000 |
| 80 %tile | 1.000 | 1.000 | 1.000 | 1.000 |
| 90 %tile | 1.000 | 1.000 | 1.000 | 1.000 |
| Maximum | 1.000 | 1.000 | 1.000 | 1.000 |
|  |  |  |  |  |
| **Reliability** |  |  |  |  |
| Mean of providers | 0.966 | 0.969 | 0.944 | 0.942 |
| Case-weighted | 0.988 | 0.988 | 0.968 | 0.973 |
| With exclusions: |  |  |  |  |
| Mean of providers | n/a | n/a | n/a | n/a |
| Case-weighted | n/a | n/a | n/a | n/a |

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

Reliability scores vary from 0.0 to 1.0, with a score of zero indicating that all variation is attributable to measurement error (noise, or variation across patients within providers) where as a reliability of 1.0 implies that all variation is caused by real difference in performance across accountable entities.

Reliability for the process Perioperative Temperature Management measure is consistently greater than 0.9, and thus can be considered to be very good. This reflects the inclusion of that measure in public reporting programs, the number of years that the measure has been reported and the number of cases available to test and analyze. Over time, performance has improved while reliability has remained relatively stable.

Reliability for the outcome Perioperative Temperature Management measure, as extracted from cases where a temperature was recorded and submitted to NACOR, reflects the limited number of cases and providers available to analyze. As cases and providers increase, as noted in 2a2.3, the reliability of the measure is expected to increase as well. We expect, should this measure be endorsed by NQF and added to quality reporting programs, that the increase of provider data submission will impact the performance scores and improve measure 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)*

Face validity of the measure score as an indicator of quality was systematically assessed as follows. After the measure was fully specified, a group of experts was assembled to rate face validity. The experts included 23 physicians.

We provided the detailed measure specifications to the experts and asked them to rate their agreement with the following statement: The scores obtained from the measure as specified will provide an accurate reflection of quality and can be used to distinguish good from poor quality.

The rating scale had five levels (1-5) with the following narrative anchors:

1 = Disagree; 3 = Moderate Agreement; 5 = Agree

As additional data and information become available on this measure, ASA intends to conduct further measure validity testing on this measure.

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

The results of the assessment of face validity indicate that an independent group of experts (different from those who advised on measure development) had high levels of agreement with the statement: “The scores obtained from the measure as specified will provide an accurate reflection of quality and can be used to distinguish good and poor quality.”

Mean rating = (3.78 out of 5)

|  |  |
| --- | --- |
| **Rating Scale** | **Number Who Selected this Rating** |
| 1 – Strongly Disagree | 1 |
| 2 – Disagree | 3 |
| 3 – Neither | 3 |
| 4 – Agree | 9 |
| 5 – Strongly Agree | 7 |
| Total | 23 |

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

This measure was examined through a group of experts. Out of the 23 participants, 16 agreed that the scores from the measure as specified would provide an accurate reflection of quality and 4 disagreed.

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

**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*)

Exclusions for this measure include cardiopulmonary bypass, regional nerve block and monitored anesthesia care. For 2013 data, we conducted a limited number of exclusion testing on the outcome measure where the provider had more than one case. When adding and the testing the excluded case, this resulted in no change to the number of providers, an increase of 315 total cases and 947 more provider-cases (multiple providers on one case). Mean performance score of the outcome measure declined from 0.975 to 0.965. Reliability declined from 0.644 to 0.531.

**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*)

When adding and the testing the excluded case, this resulted in no change to the number of providers, an increase of 315 total cases and 947 more provider-cases (multiple providers on one case). Mean performance score of the outcome measure declined from 0.975 to 0.965. Reliability declined from 0.644 to 0.531.

When adding the excluded cases to the measure, reliability declined. This decline in reliability supports the exclusion of these cases for this measure.

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

When adding the excluded cases to the measure, reliability declined. This decline in reliability supports the exclusion of these cases for this measure.

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

**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,** Click here to enter description

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

**2b4.3. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors 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 and not related to disparities*)

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

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

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

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

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

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

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

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

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

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

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

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

Our limited sample size for this outcome measure precludes us from providing accurate data to address this particular question.

**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*)

Our limited sample size for this outcome measure precludes us from providing accurate data to address this particular question.

**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?*)

Our limited sample size for this outcome measure precludes us from providing accurate data to address this particular question.

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

**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 criterion is directed 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).* ***If comparability is not demonstrated, the different specifications should be submitted as separate measures.***

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

Not applicable.

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

Not applicable.

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

Not applicable.

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

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

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

The measure is not an eMeasure, Composite or PRO-PM.

**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*)

The measure is not an eMeasure, Composite or PRO-PM.

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

The measure is not an eMeasure, Composite or PRO-PM.

1. Totals represent unique practices, facilities and providers over the four year period and not simply the sum of individual years. [↑](#footnote-ref-1)
2. For both the outcome and process measure, these tables present results of the analytic sample of physicians having two or more cases. [↑](#footnote-ref-2)
3. For both the outcome and process measure, these tables present results of the analytic sample of physicians having two or more cases. [↑](#footnote-ref-3)