**National Quality Forum—Measure Testing (subcriteria 2a2, 2b1-2b6)**

**Measure Number** (*if previously endorsed*)**:** 1424

**Measure Title**: Monthly Hemoglobin Measurement for Pediatric Patients

**Date of Submission**: 1/7/2019

**Type of Measure:**

|  |  |
| --- | --- |
| Outcome (*including PRO-PM*) | Composite – ***STOP – use composite testing form*** |
| Intermediate Clinical Outcome | Cost/resource |
| Process *(including Appropriate Use)* | 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, 2b1, 2b2, and 2b4 must be completed.** * **For outcome and resource use measures**, section **2b3** also must be completed. * If specified for **multiple data sources/sets of specificaitons** (e.g., claims and EHRs), section **2b5** 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 (2b1-2b6) 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 25 pages (*incuding questions/instructions;* minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.*** * Contact NQF staff regarding questions. Check for resources at [Submitting Standards webpage](http://www.qualityforum.org/Measuring_Performance/Submitting_Standards.aspx). * For information on the most updated guidance on how to address social risk factors variables and testing in this form refer to the release notes for version 7.1 of the Measure Testing Attachment. |

|  |
| --- |
| **Note:** The information provided in this form is intended to aid the Standing Committee and other stakeholders in understanding to what degree the testing results for this measure meet NQF’s evaluation criteria for testing.  **2a2.** **Reliability testing** [**10**](#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 **instrument-based measures** (including PRO-PMs) **and composite performance measures**, reliability should be demonstrated for the computed performance score.  **2b1.** **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 **instrument-based measures (including PRO-PMs) and composite performance measures**, validity should be demonstrated for the computed performance score.    **2b2.** **Exclusions** are supported by the clinical evidence and are of sufficient frequency to warrant inclusion in the specifications of the measure; [**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)  **2b3.** **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 social risk 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.   **2b4.** 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.  **2b5.** **If multiple data sources/methods are specified, there is demonstration they produce comparable results**.  **2b6.** 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. The degree of consensus and any areas of disagreement must be provided/discussed.  **12.** Examples of evidence that an exclusion distorts measure results include, but are not limited to: frequency of occurrence, variability of exclusions across providers, and sensitivity analyses with and without the exclusion.  **13.** Patient preference is not a clinical exception to eligibility and can be influenced by provider interventions.  **14.** Risk factors that influence outcomes should not be specified as exclusions.  **15.** With large enough sample sizes, small differences that are statistically significant may or may not be practically or clinically meaningful. The substantive question may be, for example, whether a statistically significant difference of one percentage point in the percentage of patients who received smoking cessation counseling (e.g., 74 percent v. 75 percent) is clinically meaningful; or whether a statistically significant difference of $25 in cost for an episode of care (e.g., $5,000 v. $5,025) is practically meaningful. Measures with overall less-than-optimal performance may not demonstrate much variability across providers. |

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

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

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

|  |  |
| --- | --- |
| **Measure Specified to Use Data From:**  **(*must be consistent with data sources entered in S.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: 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*).

CROWNWeb

For the Spring 2019 Maintenance submission, CROWNWeb and Medicare claims data was used.

**1.3. What are the dates of the data used in testing**? January 2013 to December 2013

For the Spring 2019 Maintenance submission, data from January – December 2017 was used.

**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: 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*)

Fifty-nine facilities that had at least 11 eligible pediatric patients during January 2013 – December 2013 were included in the analyses. Public reporting of this measure on DFC or in the ESRD QIP would be restricted to facilities with at least 11 eligible patients for the measure. We have applied this restriction to all the reliability and validity testing reported here.

For the 2019 Maintenance submission, 62 facilities that had at least 11 eligible pediatric patients during January 2017 – December 2017 were included in the analyses.

Public reporting of this measure on DFC or in the ESRD QIP would be restricted to facilities with at least 11 eligible patients for the measure to comply with restrictions on reporting of potentially patient identifiable information related to small cell size. We have applied this restriction to all the reliability and validity testing reported here.

**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*)

There were a total of 1,978 eligible pediatric patients. Among those patients, the average age was 10.5 years, 56% of patients were male, 63% were white, 28% were black, 5% were Asian/Pacific Islander, 29% were Hispanic, and 26% had glomerulonephritis as the primary cause of ESRD.

A subset of 1,280 pediatric patients who belonged to the facilities that had at least 11 eligible pediatric patients were included in the testing and analyses.

For the Spring 2019 Maintenance submission, there were a total of 1,924 eligible pediatric patients. Among those patients, the average age was 10.3 years, 58% of patients were male, 66% were white, 27% were black, 5% were Asian/Pacific Islander, 28% were Hispanic, and 31% had glomerulonephritis as the primary cause of ESRD.

A subset of 1,415 pediatric patients who belonged to the facilities that had at least 11 eligible pediatric patients were included in the testing and analyses. Among those patients, the average age was 9.8 years, 57% of patients were male, 68% were white, 25% were black, 5% were Asian/Pacific Islander, 31% were Hispanic, and 30 % had glomerulonephritis as the primary cause of ESRD.

**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**.

N/A

**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.

N/A

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

**2a2. RELIABILITY TESTING**

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

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

We used January 2013 – December 2013 CROWNWeb data to calculate facility level monthly and annual performance scores. Fifty-nine facilities that had at least 11 eligible patients were included in the testing. There were a total of 1,280 patients.

We assessed reliability by calculating facility-level Pearson correlation coefficients between the current performance month and the preceding month for reporting months during January 2013 – December 2013.

In addition, we calculated inter-unit reliability (IUR) for each reporting month and the overall 12 months. The monthly based measure was a simple average across individuals in the facility. The NQF-recommended approach for determining measure reliability is a one-way analysis of variance (ANOVA), in which the between and within facility variation in the measure is determined. The inter-unit reliability (IUR) measures the proportion of the measure variability that is attributable to the between-facility variance. The yearly based measure, however, is not a simple average and we instead estimate the IUR using a bootstrap approach, which uses a resampling scheme to estimate the within facility variation that cannot be directly estimated by ANOVA. We note that the method for calculating the IUR was developed for measures that are approximately normally distributed across facilities. Since this measure is not normally distributed, the IUR value should be interpreted with some caution.

For the Spring 2019 Maintenance submission, we used January 2017 – December 2017 CROWNWeb data to calculate facility level monthly and annual performance scores. Sixty-two facilities that had at least 11 eligible patients were included in the testing. There were a total of 1,415 patients.

We assessed reliability by calculating facility-level Pearson correlation coefficients between the current performance month and the preceding month for reporting months during January 2017 – December 2017. We also calculated inter-unit reliability using the same methodology as the previous submission.

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

The Pearson correlation coefficients of each pair of the current and the preceding months ranged from 0.78 to 0.98. All were statistically significant (p<0.0001), indicating this measure is reliable over time.

For reliability we calculated the monthly and annual IUR across the 12 reporting months. As explained above, the method for calculating the IUR was developed for measures that are approximately normally distributed across facilities.  IUR=0.96, which is high and suggests 96% of variation in the measure is attributed to between facility variation.

For the Spring 2019 Maintenance submission, the median of Pearson correlation coefficients of each pair of the current and the preceding months was 0.84, with a range of 0.28 to 0.92. All were statistically significant (p<0.05), indicating this measure is reliable over time.

In a sensitivity analysis, we identified an outlier facility that drastically influenced the correlation between February and March. After removing the outlier facility, the range of correlation coefficients of each pair of the current and the preceding months narrowed down to be from 0.47 and 0.92, with a median of 0.85. All correlations were statistically significant (p<0.0001),

For reliability we calculated the monthly and annual IUR across the 12 reporting months. As explained above, the method for calculating the IUR was developed for measures that are approximately normally distributed across facilities.  IUR=0.82, which is high and suggests 82% of variation in the measure is attributed to between facility variation.

**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 Pearson correlation coefficients were moderate to strong, indicating this measure is reliable over time periods of measurement.

The IUR suggest this measure is reliable. However, since the distribution of performance scores is skewed, the IUR value should be interpreted with some caution.

For the Spring 2019 Maintenance submission, the Pearson correlation coefficients were moderate to strong, indicating this measure is reliable over time periods of measurement.

The IUR again suggests this measure is reliable. However, since the distribution of performance scores is skewed, the IUR value should be interpreted with some caution.

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

**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)*

We used January 2013 – December 2013 CROWNWeb data to calculate facility level monthly and annual performance scores. Fifty-nine facilities that had at least 11 eligible patients and included 1,280 patients in total were included in the testing.

We computed the Spearman correlation to assess the association between the annual performance scores and the NQF endorsed (0369) standardized mortality ratio (SMR) using the 2013 SMR. The data source and the methodology in SMR calculations are attached.

This measure is being maintained on the basis of face validity. The measurement of hemoglobin as a dialysis quality measure was initially developed and approved by a Clinical TEP, which agreed that this quality measure is important in the assessment of the quality of care for pediatric dialysis patients.

For the Spring 2019 Maintenance submission, we used January 2017 – December 2017 CROWNWeb data to calculate facility level monthly and annual performance scores. Sixty-two facilities that had at least 11 eligible patients and included 1,415 patients in total were included in the testing.

We computed the Spearman correlation to assess the association between the annual performance scores and the NQF endorsed (0369) standardized mortality ratio (SMR) using the 2017 SMR. We expect the correlation to be negative, insofar as facilities with successful processes for monitoring clinically important intermediate outcomes of care would be expected to have better primary outcomes, including lower mortality. In addition, if consistent, effective monitoring of hemoglobin outcomes contributes to a higher percentage of hemoglobin values in the facility’s target range, then patient primary outcomes would likely be improved (lower mortality), assuming the facility’s target ranges were appropriate.

This measure is being maintained on the basis of face validity. The measurement of hemoglobin as a dialysis quality measure was initially developed and approved by a Clinical TEP, which agreed that this quality measure is important in the assessment of the quality of care for pediatric dialysis patients.

**2b1.3. What were the statistical results from validity testing**? (*e.g., correlation; t-test*)

Spearman correlation coefficient was -0.20, p=0.13.

For the Spring 2019 Maintenance submission, the spearman correlation coefficient was 0.07, p=0.55.

**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 result suggests that facilities with a higher percentage of pediatric patients (calculated as patient months) with hemoglobin measured is associated with a lower risk of mortality relative to facilities with a lower percentage of pediatric patients with hemoglobin measured. The result is however not statistically significant. This may be due to the small sample size (e.g., small number of facilities eligible for this measure).

The result does not suggest a statistically significant association between the measure (calculated as patient months) and mortality. We maintain the measure based on face validity as determined by the Technical Expert Panel that initially developed the measure.

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

**2b2. EXCLUSIONS ANALYSIS**

**NA**  **no exclusions — *skip to section*** [***2b3***](#section2b4)

**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*)

N/A

**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*)  
N/A

**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*)  
N/A

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

**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*** [***2b4***](#section2b5)***.***

**2b3.1. 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

**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.**

N/A

**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**.   
Risk adjustment is not required for this measure. While trend analysis suggests by disparities by race, the conservative interpretation would be that these differences reflected in the trend analysis reflect disparities in care for certain subpopulations. In the absence of biological effects explaining these differences, risk adjustment for these factors would potentially mask disparities in care.

**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?  
N/A

**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?**N/A

**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.**

N/A

**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***](#question2b49)

N/A

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

N/A

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

**2b3.8. Statistical Risk Model Calibration – Risk decile plots or calibration curves**:  
N/A

**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*)

N/A

**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*)

N/A

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

**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)*

Given that the number of facilities included in the calculation in 1.5 is small, the sample was determined to be too small to display useful data on meaningful differences.

For the Spring 2019 submission, we followed the methodology described below to using data from January – December 2017.

Testing was performed on the yearly based performance score. We used a re-sampling based exact test, with re-sampling generated from the population distribution of the patient level outcomes. Note that a one-sided test with significance level 0.025 is used (corresponding to cutoff=0.05 in a two-sided test) due to non-symmetric structure of the measure's distribution. To calculate the p-value, we compute the probability that the facility would experience a number of events (i.e., percentage with hypercalcemia) more extreme than that observed if the null hypothesis were true, with the null hypothesis being that the facility's distribution of hypercalcemia will follow the overall national distribution.

**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*)

N/A

For the Spring 2019 Maintenance submission, we did not identify statistically significant in performance scores across 62 facilities that had at least 11 eligible patients. This is because of the size of patient population is relatively small within the each facility (mean=23, median=20) and the performance score of measure is distributed uniformly across facilities (mean= 90.2%, standard deviation=11%).

**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?*)

N/A

For the Spring 2019 maintenance submission, as described above we did not identify statistically significant differences in performance scores for this measure, due to the small patient population in each facility.

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

**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*)

N/A

**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*)

N/A

**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*)

N/A

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

**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 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*)

Missing is the outcome and this measure is reporting the percentage of non-missing. Thus, the missing data does not cause bias in this measure.

**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*)

N/A

**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 nonresponders) and how the specified handling of missing data minimizes bias**?** (i*.e., what do the results mean in terms of supporting the selected approach for missing data and what are the norms for the test conducted; if no empirical analysis, provide rationale for the selected approach for missing data*)

N/A