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

**Measure Number** (*if previously endorsed*)**:** 1425

**Measure Title**: Measurement of nPCR for Pediatric Hemodialysis 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 and Medicare Claims Data from January 2013 to December 2013

For the Spring 2019 maintenance submission, CROWNWeb and Medicare Claims Data from January 2017 to December 2017.

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

For the Spring 2019 maintenance submission, January 2017 to December 2017

**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*)

The measured entities used in testing and analysis include reported nPCR and the necessary data elements needed for calculating nPCR for 455 in-center hemodialysis (ICH) pediatric patients from 30 dialysis facilities with at least 11 eligible pediatric patients across all regions of the United States.

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.

Facilities vary in size, and include anywhere from 11 to 28 eligible ICH pediatric patients. The data elements include “nPCR” or the combination of “Kt/V hemodialysis collection date”, “BUN pre-dialysis”, “BUN post-dialysis”, “pre-dialysis weight”, “pre-dialysis weight unit of measure”, “post-dialysis weight”, “post-dialysis weight unit of measure”, “delivered minutes of BUN hemodialysis session”, and “interdialytic time.”

For the Spring 2019 maintenance submission, the measured entities used in testing and analysis include reported nPCR and the necessary data elements needed for calculation of nPCR. There are 511 in-center hemodialysis (ICH) pediatric patients from 29 dialysis facilities that have had at least 11 eligible pediatric patients across all regions of the United States.

Public reporting of this measure on DFC or in the ESRD QIP would be restricted to facilities with at least 11 eligible patients in order 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.

Facilities vary in size, and include anywhere from 11 to 42 eligible ICH pediatric patients.

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

Testing was performed on all Medicare and non-Medicare pediatric, ICH patients available in CROWNWeb from 2013. The sample included 455 patients from 225 facilities. The table below shows the number and percent of pediatric ICH patients by race, sex, and Hispanic ethnicity.

|  |  |  |
| --- | --- | --- |
| **Race** | **Frequency** | **Percent** |
|  |  |  |
| Asian | 23 | 5.05% |
| Black | 147 | 32.31% |
| White | 274 | 60.22% |
| Native American | 5 | 1.10% |
| Pacific Islander | 4 | 0.88% |
| Mid East Arabian | 1 | 0.22% |
| Other/Multi-racial | 1 | 0.22% |
|  |  |  |
| **Sex** |  |  |
| Female | 202 | 44.40% |
| Male | 253 | 55.60% |
|  |  |  |
| **Ethnicity** |  |  |
| Hispanic | 163 | 35.82% |
| Non-Hispanic | 292 | 64.18% |

For the Spring 2019 maintenance submission, testing was performed on all Medicare and non-Medicare pediatric, ICH patients available in CROWNWeb from 2017. The sample included 511 patients from 29 facilities. The table below shows the number and percent of pediatric ICH patients by race, sex, and Hispanic ethnicity.

|  |  |  |
| --- | --- | --- |
| **Race** | **Frequency** | **Percent** |
|  |  |  |
| Asian/Pacific Islander | 37 | 7.24% |
| Black | 146 | 28.57% |
| White | 317 | 62.04% |
| American Indian/Alaskan Native | 3 | 0.59% |
| Other/Multi-racial | 8 | 1.57% |
|  |  |  |
| **Sex** |  |  |
| Female | 237 | 46.38% |
| Male | 274 | 53.62% |
|  |  |  |
| **Ethnicity** |  |  |
| Hispanic | 169 | 33.07% |
| Non-Hispanic | 339 | 66.34% |
| \*\*3 missing |  |  |

**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*)

January 2013 – December 2013 CROWNWeb data were used to calculate the inter-unit reliability (IUR) for the overall 12 months to assess the reliability of this measure. 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 IUR was estimated 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 followed the same methodology as that used in the previous submission for the data from January 2017 – December 2017

**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 overall IUR was 0.985, which indicates that about 98.5% of the variation in the measure can be attributed to the between facility differences and 1.5% to within facility variation.

For the Spring 2019 maintenance submission, the overall IUR was 0.963, which indicates that about 96.3% of the variation in the measure can be attributed to the between facility differences and 3.7% to the within 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 IUR suggests 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 IUR suggests that 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)*

Concurrent validity was used as a method for testing the association between facility percentage of reporting nPCR month and mean nPCR value. Using calendar year 2013 CROWNWeb data, average facility-mean nPCR value was compared between the two groups using a two-sided two-sample t-test. Facilities were then categorized into one of two groups:

1) Facilities with 100% reporting of nPCR among their pediatric patients;

2) Facilities with less than 100% reporting of nPCR among their pediatric patients

nPCR values outside the range of [0.2, 1.8] were excluded.

This measure was also reviewed and approved by a Clinical TEP in 2010.

For the Spring 2019 maintenance submission, we employed methods similar to those used in the previous submission. The current analysis is based on January 2017 – December 2017 data, and employed the following categorization:

1. Facilities with >= 85% reporting of nPCR among their pediatric patients;
2. Facilities with <85% reporting of nPCR among their pediatric patients.

We performed a validity analysis to examine the association between facility percentage of reporting nPCR month and mean nPCR value via the means of two-sample t-test. We would expect that facilities with at least 85% reporting of nPCR among their pediatric patients are likely paying attention to this parameter in their clinical management (i.e., assessment of protein intake) of pediatric dialysis patients.

We also maintain this measure on the basis of face validity, as the measure was reviewed and approved by a Clinical TEP in 2010.

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

Among facilities with at least 11 eligible pediatric patients with recorded nPCR values, facilities with 100% reporting of recorded nPCR values had a mean serum albumin of 3.77, while facilities with less than 100% reporting of recorded nPCR values had a mean serum albumin of 4.0. Using a t-test, these values were statistically significant (p-value 0.02).

For the Spring 2019 maintenance submission, among facilities with at least 11 eligible pediatric patients and recorded nPCR values, facilities with 85% or higher reporting of recorded nPCR values had a mean nPCR of 0.9974, while facilities with less than 85% reporting of recorded nPCR values had a mean nPCR of 0.6587. According to the t-test (Satterthwaite version), the mean nPCR values of these two groups were not statistically significant (p-value=0.13)

**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?*)

These findings are somewhat unexpected, and in the opposite direction of analyses previously conducted. This difference may have resulted from a larger sample utilized for the current analyses (previous analyses were conducted over a limited timeframe). We speculate that the observed findings may have resulted if facilities are more likely to collect necessary data elements for nPCR assessment in patients for which nutritional concerns exist. These results therefore do not necessarily contradict the importance of evaluating nPCR.

Give no evidence that facility-specific nPCR differs by reporting percentage, we found no evidence of invalidity.

In addition to these results, along with the clinical importance of evaluating nPCR, we also propose maintaining 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**.   
N/A

**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)*

Differences in measure performance were evaluated separately for each facility using patient level analyses. The proportion of patients with yearly based percent of patients with reporting of nPCR was compared between one facility and the overall national distribution, and repeated for each individual facility.

Note that the monthly based measure is a simple average of binary outcomes across individuals in the facility, for which the binary outcome equals to 0 (failure = fail to report nPCR) if the value is missing. The differences in proportions can be compared using Fisher’s Exact tests or its normal approximation. The yearly based measure, however, is not a simple average of binary outcomes and we instead used a re-sampling based exact test, with re-sampling generated from the population distribution of the patient level outcomes. More details for the testing method are provided in Appendix. Due to non-symmetric of the measure distributions, one-sided test with significance level 0.025 is used (corresponding to cutoff=0.05 in two-sided test). To calculate the p-value, we assess the probability that the facility would experience a number of events more extreme than that observed if the null hypothesis were true.

For the Spring 2019 submission, we reproduced the significance analysis using data from January – December 2017. We have revised the description of the analysis to be clearer:

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*)

Proportion of facilities with significant p-values (0-as expected/better than expected; 1-worse than expected; cutoff=0.025) is shown as follows:

|  |  |  |  |
| --- | --- | --- | --- |
|  | # of Facilities | Percent of facilities | Median Performance Score |
| As Expected/Better than Expected | 23 | 76.67% | 100.00% |
| Worse than Expected | 7 | 23.33% | 24.49% |

For the Spring 2019 maintenance submission, the proportion of facilities with significant p-values (0-as expected/better than expected; 1-worse than expected; cutoff=0.025) is shown as follows:

|  |  |  |
| --- | --- | --- |
|  | # of Facilities | Percent of facilities |
| As Expected/Better than Expected | 27 | 93.1% |
| Worse than Expected | 2 | 6.9% |

**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?*)

Significance testing identifies 7 facilities (23.3%) with worse than expected performance at a median of 24.5% of patients with reporting of nPCR data elements. The clear separation in measure performance between facilities identified with worse than expected performance versus those with as expected or better than expected performance provides support for the ability to identify clinically important differences in performance on this measure through significance testing.

For the Spring 2019 submission, significance testing identifies that 27 facilities (93.1%) have achieved expected performance, and 2 facilities (6.9%) had worse than expected performance. Between facilities identified with worse than expected performance versus those with as expected or better than expected performance, there exists a clear separation that provides support for the ability to identify clinically important differences in performance on this measure through significance testing.

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

**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