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

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

**Measure Title**: Ultrafiltration rate greater than 13 ml/kg/hr

**Date of Submission**: 9/25/2014

**Type of Measure:** Intermediate outcome

|  |  |
| --- | --- |
| 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*).

2013 CROWNWeb data

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

The measured entities include ultrafiltration rate values (calculated from non-missing data elements in pre-dialysis weight (kg), post-dialysis weight (kg), and delivered time of dialysis (mins)) from 400,308 chronic adult ESRD patients on hemodialysis from 5,556 dialysis facilities with a minimum of 11 patients across all regions of the United States in 2013. Facilities vary in size, and include anywhere from 11 to 448 patients.

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.

**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 adult (>=18 years old) hemodialysis patients at dialysis facilities with a minimum of 11 patients for 2013 using CROWNWeb data. The sample included 400,308 patients from 5,556 facilities.

|  |  |  |
| --- | --- | --- |
| **Race** | **Frequency** | **Percent** |
| White | 160384 | 54.03% |
| Black | 116165 | 39.13% |
| Asian/Pacific Islander | 14153 | 4.77% |
| Native American | 4371 | 1.47% |
| Mid East Arabian  /Indian Subcontinent  /Multi-racial/other | 1761 | 0.59% |
|  |  |  |
| **Sex** |  |  |
| Female | 129780 | 43.72% |
| Male | 167025 | 56.27% |
|  |  |  |
| **Ethnicity** |  |  |
| Hispanic | 54088 | 18.22% |
| Non-Hispanic | 242705 | 81.78% |
|  |  |  |
| **Age** |  |  |
| 18 to 64 | 160836 | 54.18% |
| 65 and Older | 135999 | 45.82% |

**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**.

Validity testing was against 2013 outcomes. CROWNweb 2013 data consisting of 400,308 patients from 5,556 facilities was merged with 2013 facility-level outcomes data used for the 2014 Dialysis Facility Report (DFR), and was restricted to facilities with at least 11 patients. The merged file consisted of 3,566 facilities. The sample used for validity testing was also used for a risk-adjusted analysis in section 2b4.2.

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

**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.

**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 0.84.

**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 of 0.84indicates that about 84% of the variation in the UFR>13 can be attributed to the between facility differences and 16% to within facility variation. This suggests that most of the variation observed in the measure is due to the differences between facilities as opposed to random variation within facilities.

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

**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)*

Using CROWNWeb data for 2013, a Poisson regression analysis was performed with observed hospitalization, and mortality, respectively, by quintile levels of facilities with ultrafiltration > 13 ml/kg/hr. The model using mortality was offset by expected mortality. The models included only facilities with at least 11 patients. The relative risk and 95% confidence intervals for hospitalization, and mortality, respectively, were reported for UFR quintiles 2-5, in reference to quintile 1. These were unpublished internal analyses performed by the University of Michigan - Kidney Epidemiology and Cost Center.

In addition, in June 2013 there was an assessment of face validity based on polling of a CMS Technical Expert Panel (TEP) that took place after the in-person conference held in April 2013. 5/8 voting members of the Technical Expert Panel (TEP) voted to recommend development of a facility-level measure for reporting the percent of patients at dialysis facilities with an ultrafiltration rate greater than 13 ml/kg/hr. A recent observational study that examined different thresholds of UFR and their relationship to cardiovascular mortality showed an increased risk of cardiovascular mortality for UFR greater than 13 ml/kg/hr.

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

*Hospitalizations (admissions):*

|  |
| --- |
| ***Poisson Regression of Facility UFR > 13 Quintiles*** |
| ***Facilities with at least 11 patients and at least 5 patient years at risk*** |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Parameter** | **N** | **Estimate** | **Standard Error** | **Relative  Risk** | **Wald 95% Confidence Limits** | | **Wald Chi-Square** | **Pr > ChiSq** |
| **Intercept** |  | -0.0537 | 0.0086 |  | -0.071 | -0.037 | 39.05 | <.0001 |
| **Reference (Quintile 1) (0 - 9.49)** | 1111 |  |  |  |  |  |  |  |
| **Quintile 2 (9.50 - 13.18)** | 1111 | 0.0489 | 0.0116 | **1.05** | 0.0262 | 0.0717 | 17.74 | **<.0001** |
| **Quintile 3 (13.19 - 16.78)** | 1112 | 0.0731 | 0.0115 | **1.08** | 0.0506 | 0.0957 | 40.4 | **<.0001** |
| **Quintile 4 (16.78 - 21.61)** | 1110 | 0.061 | 0.0116 | **1.06** | 0.0383 | 0.0837 | 27.71 | **<.0001** |
| **Quintile 5 (21.62 - 74.19)** | 1112 | 0.0286 | 0.0117 | **1.03** | 0.0057 | 0.0515 | 6.01 | **0.0143** |

*Mortality:*

|  |
| --- |
| ***Poisson Regression of Facility UFR > 13 Quintiles*** |
| ***Facilities with at least 11 patients and at least 3 expected deaths*** |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Parameter** | **N** | **Estimate** | **Standard Error** | **Relative Risk** | **Wald 95% Confidence Limits** | | **Wald Chi-Square** | **Pr > ChiSq** |
| **Intercept** |  | -0.0327 | 0.011 |  | -0.054 | -0.011 | 8.87 | 0.0029 |
| **Reference (Quintile 1) (0 - 9.49)** | 1111 |  |  |  |  |  |  |  |
| **Quintile 2 (9.50 - 13.18)** | 1111 | 0.0383 | 0.0149 | **1.04** | 0.009 | 0.0676 | 6.55 | **0.0105** |
| **Quintile 3 (13.19 - 16.78)** | 1112 | 0.0408 | 0.0149 | **1.04** | 0.0116 | 0.07 | 7.48 | **0.0062** |
| **Quintile 4 (16.78 - 21.61)** | 1110 | 0.0599 | 0.0149 | **1.06** | 0.0306 | 0.0892 | 16.08 | **<.0001** |
| **Quintile 5 (21.62 - 74.19)** | 1112 | 0.0287 | 0.0151 | 1.03 | -9E-04 | 0.0583 | 3.61 | 0.0573 |

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

The analyses presented above examine associations between quintiles of facility performance (defined by the percent of patients with an ultrafiltration rate above 13 ml/kg/hr) and the outcomes of hospitalization and mortality. In general, for both outcomes, facilities in the lowest quintile (i.e. who had the least percent of patients with an ultrafiltration rate above 13 ml/kg/hr) had a significantly lower risk of hospitalization or mortality compared to higher quintiles. This provides support for the potential beneficial impact of a measure to encourage facilities to limit ultrafiltration rates in their patients. Of note, there was no clear gradient of risk across increasing quintiles, driven by the more modest risk in the highest quintile (which also just failed to achieve statistical significance for the outcome of mortality). This is likely a manifestation of selection bias relating to healthier patients being more tolerant of higher ultrafiltration rates. Facilities in the highest quintile may therefore have greater proportion of healthier patients in their panel, blunting the apparent risk of higher ultrafiltration rates. Adjustment for this effect would require complex analyses beyond the scope of what is possible here. However, previously published patient-level analyses have consistently demonstrated worse outcomes associated with higher ultrafiltration rates (Movilli 2007, Flythe 2011, Saran 2006, Flythe 2013, Saran 2006, Shoji 2004, Burton 2009, McIntyre 2010)\*

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

**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*)

N/A

**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*)

N/A

**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*)

N/A

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

**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**.

No risk adjustment or stratification is being performed at this time. The impact of a risk-adjusted measure, especially in terms of body size (i.e. BMI, BSA), is not supported in the current body of evidence for UFR>13 ml/kg/hr. The UFR>13 measure is strictly indexing to its relevance as a measure of rate of fluid removal. While noticeable differences in the percentages of UFR>13 were observed in various subgroups at the national level, stratification at the facility level would result in few facilities with sufficient subgroup populations for stratified reporting. These differences will continue to be monitored in aggregated data, such as at the national level (see table in appendix for 1b.4).

**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*)

N/A

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

N/A

**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*)

N/A

*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*)**:**

N/A

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

N/A

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

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

N/A

**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*)

N/A

**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*)

N/A

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

**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)*

Differences in measure performance were evaluated separately for each facility using patient level analyses. The proportion of patients with yearly based percentage of patients with UFR>13 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) if the value is less than the threshold or 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.

**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*)

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

Classification of facility performance for UFR > 13 (CROWNWeb 2013)

|  |  |  |  |
| --- | --- | --- | --- |
|  | # of Facilities | Percent of facilities | Median Performance Score |
| As Expected/Better than Expected | 4656 | 83.80% | 13.50% |
| Worse than Expected | 900 | 16.20% | 26.60% |

**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?*)

Significance testing identifies 900 facilities (16.2%) with worse than expected performance at a median of 26.6% of patients with UFR >13. 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.

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

**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*)

N/A

**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*)

N/A

**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*)

N/A

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

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

N/A

**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*)

N/A

**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*)

N/A

**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*)

N/A