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

**Measure Number** (*if previously endorsed*)**: Not applicable; new measure**

**Measure Title**: **Post-Dialysis Weight Above or Below Target Weight**

**Date of Submission**: 2/26/2015

**Type of Measure:**

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

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

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

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

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

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

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

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

This KCQA performance measure was tested using data from three KCQA member dialysis organizations, each with the capacity to provide retrospective analyses from a data warehouse/repository. All pertinent data from all eligible (i.e., adult in-center hemodialysis) patients of the participating organizations during the testing period were included in the datasets.

**1.3. What are the dates of the data used in testing**? January 1, 2013-December 31, 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: Dialysis facility | other: Dialysis facility |

**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 entity is the dialysis facility. The following is a composite description of the facility demographics across the three dialysis organizations that participated in testing:

* Testing encompassed 4,252 dialysis facilities.
* Based on monthly averages, the mean facility census (i.e., the number of patients receiving care at the facility), weighted and averaged across the three organizations, was 84.11 patients.
* Facility census ranged from 1 to 664 patients per month.

Facility demographics for each of the three dialysis organizations is summarized in the following table:\*

|  |  |  |  |
| --- | --- | --- | --- |
| **DIALYSIS ORGANIZATION** | **NUMBER OF FACILITIES IN STUDY** | **MEAN FACILITY SIZE** | **RANGE OF FACILITY SIZES** |
| **A** | 212 | 54.81 patients | 1-188 patients |
| **B** | 2,047 | 64.25 patients | 1-487 patients |
| **C** | 1,993 | 113.91 patients | 1-664 patients |

\* To preserve anonymity, data are presented as coming from Organization A, B, and C. This nomenclature is random and is scrambled throughout the measure submission documents such that Organization A in one section might become Organization B or C in another section.

**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*)

412,522 patients across the three organizations met the measure’s denominator criteria and were included, with a range of 15,184 to 215,008 patients per organization. The following is a composite description of patient demographics:

* Mean patient age = 61.66 years
* Range of patient ages = 18.01 to 104.00 years
* Gender = 56.26% male and 43.74% female
* Race/Ethnicity = 52.37% white, 36.33% African American, 2.82% Asian, 1.16% American Indian/Native Alaska, 0.67% Native Hawaiian/other Pacific Islander, 0.57% other/missing/declined; 15.60% Hispanic (independent of race)

Patient demographics for each organization is summarized in the following table:\*

|  |  |  |  |
| --- | --- | --- | --- |
|  | **A** | **B** | **C** |
| **TOTAL NUMBER PATIENTS** | 182,330 | 215,008 | 15,184 |
| **RANGE OF PATIENT AGES** | 18.01-84.93 years | 18.01-104.00 years | 18.01-100.01 years |
| **MEAN PATIENT AGE** | 62.24 years | 61.15 years | 61.90 years |
| **GENDER: MALE** | 56.30% | 56.25% | 56.27% |
| **GENDER: FEMALE** | 43.70% | 43.75% | 43.73% |
| **RACE: WHITE** | 59.58% | 38.31% | 47.25% |
| **RACE: BLACK/ AFRICAN AMERICAN** | 36.24% | 35.89% | 43.59% |
| **RACE: AMERICAN INDIAN/ ALASKA NATIVE** | 0.87% | 1.32% | 2.39% |
| **RACE: ASIAN** | 1.93% | 3.62% | 2.21% |
| **RACE: NATIVE HAWAIIAN/ OTHER PACIFIC ISLANDER** | 0.42% | 0.91% | 0.29% |
| **RACE: OTHER/ DECLINE TO STATE/MISSING** | 0.96% | 0.09% | 2.79% |
| **ETHNICITY: HISPANIC (independent of race)** | 14.30% | 17.32% | 6.73% |

**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**.

While the original measure specifications required that all dialysis sessions for all eligible (i.e., adult in-center hemodialysis) patients be included in facilities’ score calculations, one dialysis organization also tested an “abbreviated” version of the measure, wherein only those dialysis sessions (including supplemental sessions) falling within the same week that the monthly Kt/V is drawn are included. Such an approach was tested to harmonize the measure specifications with those of another KCQA fluid management measure and to significantly reduce burden on dialysis facilities not permitted to batch submit data to CMS. This arm of testing encompassed all of that organization’s facilities and patients.

Based on the testing results, the specifications advanced to NQF specify the “abbreviated” approach—i.e., only those sessions falling within the same week that the monthly Kt/V is drawn.

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

**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*)

For both the full (all sessions) and abbreviated versions of the measure, the reliability of the measure was assessed using a repeated measures analysis of variance (ANOVA) test. Data were statistically analyzed using the facility and the treatment month as independent variables and the measure scores as dependent variables. Analyses were conducted separately for all the facilities in each of the three dialysis organizations. The rationale for using a repeated measures ANOVA is that there should be relatively little within-facility variation in the monthly proportion of patients’ dialysis treatment sessions that do not meet the clinical standard threshold (e.g., are not within +/- 1.0 kg of the difference between the observed and target post-treatment weight). Rather, if the measure is helpful in discriminating between high and low performing dialysis facilities, the level of variation from month to month should be high between facilities.

The KCQA measure was analyzed for within- and between-facility variance among patients’ dialysis sessions that did not meet the quality standard specifications. The “within” facility variation is the “error variance” or “noise” that reflects the degree of between-month variation in the measure that occurs within a facility. The “between” facility variation is the explained or “systematic variance” (i.e., the “signal”) that is attributable to variation in performance between facilities and represents real differences in performance. The intra-class correlation coefficient (ICC) was calculated to estimate the ratio of the between- to the within-facility variance, standardized for both the level of variation and the numbers of observations examined. The higher the ICC, the greater the reliability of the measures. The ratio of the between- to the within-facility variation was also examined as a “signal to noise” ratio. Greater between-facility variation than within-facility variation indicates that the measure is discriminating between 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 table below reports the measure’s intra-class correlation coefficient for each of the participating dialysis organizations, as well as the ratio of between- to within-facility variation:

|  |  |  |
| --- | --- | --- |
| **Dialysis Organization** | **Intra-Class Correlation** | **Ratio of Between- to Within-Facility Co-Variance** |
| A | 0.77 | 3.6 |
| B | 0.71 | 2.5 |
| C | 0.63 | 2.0 |

Reliability statistics for the “abbreviated” measure specifications were nearly identical to those of the “every session” construct.

**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?*)

As demonstrated above for both the “every session” and the “abbreviated” measure constructs, the intra-class correlation for all organizations is high, indicating a good level of reliability within facilities over time and considerable within-facility stability with respect to performance on this measure over the course of the 12 months. Additionally, the estimated between-facility variation is greater than the within-facility variation across all organizations, which, when considered in light of the high intra-class correlation coefficients, suggests that both versions of the measure are reliable and differentiate between 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)*

Criterion predictive (correlative) validity was used to assess the correlation of the computed measure scores against some criterion (e.g., another measure of the same construct or an outcome) deemed valid. Specifically, the validity of the measure was evaluated by correlating facility-specific scores with each facility’s 2013 Standardized Hospitalization Ratio for Admissions measure (SHR, NQF #1463) and Standardized Mortality Ratio\* measure (SMR, NQF #0369) scores, using Pearson’s Correlation Coefficient. Both the SHR and SMR are NQF-endorsed publicly available dialysis facility outcome measures that the KCQA measure could be expected to impact.

To allow for correlation with the most current SHR/SMR scores publicly available on Dialysis Facility Compare (DFC), 2013 facility data were used for testing. If available, correlation to 2013 hospitalization rates from the facilities’ DFRs also was analyzed.

Additionally, between July 2014 and February 2015, KCQA conducted an iterative assessment of face validity based on a series of conferences of the KCQA Steering Committee and the KCQA Feasibility/ Testing Workgroup, as well as repetitive polling of the full KCQA at various stages of the measure development process.

\* The SMR specifications are based on a 4-year rolling period.

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

The KCQA measure’s Pearson’s Correlation Coefficients are summarized as follows:

|  |  |  |  |
| --- | --- | --- | --- |
| **Dialysis Organization** | **2013 SHR** | **2013 SMR** | **2013 Hospitalization Rate (from DFR)** |
| **A** | 0.16  p < 0.0001 | 0.10  p < 0.0001 | 0.19  p < 0.0001 |
| **B** | 0.17  p = 0.0169 | 0.25  p = 0.0002 | 0.17  p = 0.0169 |
| **C** | 0.23  p < 0.001 | 0.19  p < 0.001 | Not available |

For the organization that testing both constructs, validity statistics for the “abbreviated” measure specifications were identical to those of the “every session” construct.

**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?*)

For both measure constructs, the correlation between the quality performance measure of the deviation between patients’ post-dialysis and prescribed target weights and the SMR and SHR is statistically significant and in the expected direction; facilities that have fewer patients deviating from their prescribed target weight post-dialysis have lower mortality and hospitalization rates. The correlation supports the hypothesized underlying measure concept—that improving fluid management in dialysis patients will reduce the ultimate adverse patient outcomes of mortality and hospitalization.

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

**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*)

For each facility across the three participating dialysis organizations, the overall number and percentages of patients meeting each exclusion criterion was recorded for each of the 12 months. The monthly and annual frequencies of the occurrence of each exclusion and the variability of the exclusions were then analyzed.

**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*)

The annual counts for the individual denominator exclusions across the three organizations were:

* Age <18 years = 32,085 patient-months (0.66%)
* Patients receiving care in a facility <30 days = 306,860 patients-months (7.58%)
* Home dialysis patients = 192,645 patient-months (5.08%)
* <7 hemodialysis treatments in the facility during the reporting month = 335,606 patient-months (7.64%)
* Patients without a completed CMS Medical Evidence Form (Form CMS-2728) in the reporting month = 32,806 patient-months (0.65%)
* Kidney transplant recipients with a functioning graft = not tested (discussed further below)

The total number of annual exclusions across the three organizations was 636,098 patient-months, with a range of 14,949 to 440,473 patient-months excluded per organization. The average monthly exclusions across the three organizations was 53,008 patients, with a range of 1 to 288 patients excluded per facility each month.

**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*)

As can be seen, with one exception (transplant recipients with functioning graft), the frequency with which the exclusions were encountered during testing is sufficient to demonstrate they are necessary to prevent unfair distortion of performance results; Table 4 of KCQA Testing Data Attachment (attached to this form) documents that the variability in their occurrence across providers also supports the need for the exclusions.

We found that the “kidney transplant recipients with a functioning graft” exclusion could not be operationalized consistently across providers during the limited time for testing. However, we have retained it in the measure specifications since it remains clinically relevant and appropriate.

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

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

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

**No risk adjustment or stratification**

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

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

**Other,** Click here to enter description

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

**2b4.3. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors used in the statistical risk model or for stratification by risk** (*e.g., potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of p<0.10; correlation of x or higher; patient factors should be present at the start of care and not related to disparities*)

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

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

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

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

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

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

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

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

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

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

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

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

Descriptive statistics for the annual performance measure scores for all tested entities (facilities) were constructed. These statistics include the mean, standard deviation and standard error, 95% confidence interval, median, mode, range of scores, and the interquartile range of scores across the measured entities. We defined “meaningful difference” as a significant spread (>20%) between minimum and maximum scores or a significant spread between median and minimum or median and maximum scores.

**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*)

Findings are summarized in the following table:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **N** | **Range of Scores** | **Mean Score** | **Median Score** | **Mode of Scores** | **Interquartile Range** |
| 4,252 facilities | 0-100% | 23.13% | 22.56% | 25.00% | 14.08 |
| SD =10.56 |
| SE = 0.16 |
| 95% CI = 22.18-23.45% |

**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?*)

Results are interpreted as showing a significant spread between both the minimum and maximum scores, as well as the median and minimum and maximum scores, indicating that the measure identifies clinically and practically meaningful differences in performance among the measured entities.

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

**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*)

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

**2b6.3. What is your interpretation of the results in terms of 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*)

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

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

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

All facilities remained in the analysis, regardless of the magnitude of missing data; however, participating dialysis organizations were instructed to remove individual dialysis sessions missing any data necessary to calculate the measure scores from the analysis. For each facility across the three participating dialysis organizations, the overall number, average, and range of dialysis sessions with missing data were recorded for each of the 12 months. The monthly and annual frequencies of sessions with missing data were then analyzed. Note, that while we did not perform a separate examination of missing data on the abbreviated measure construct, it can be inferred that the percentage of missing data during the Kt/V week is equivalent to that of the entire month; it may in fact be less, since facilities are attuned to data collection during that week.

**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*)

Findings are provided in the table that follows. In summary, 75,188 individual dialysis sessions across all three dialysis organizations were excluded from the analysis over the testing year due to missing data. The average annual number of sessions excluded per facility was 17.68.

The average monthly number of excluded treatments across the three organizations was 6,266, with a range of 1 to 323 sessions excluded per facility each month secondary to missing data. The average monthly number of sessions excluded per facility was 1.47.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | **DIALYSIS SESSIONS EXCLUDED ACROSS ALL FACILITIES**  **(TOTAL NUMBER / AVERAGE PER FACILITY / FACILITY RANGES)** | | | | | | | | | | | | **TOTAL** |
| **Jan** | **Feb** | **Mar** | **Apr** | **May** | **Jun** | **Jul** | **Aug** | **Sep** | **Oct** | **Nov** | **Dec** | **Annual** |
| **Missing/**  **incomplete data** | **Org A** | T=1,984  A=9.36  R=(0-323) | 1,679  7.92  (0-288) | 1,853  8.74  (0-209) | 1,701  8.02  (0-152) | 3,572  16.85  (0-214) | 1,437  6.78  (0-113) | 1,554  7.33  (0-146) | 1,638  7.73  (0-156) | 1,517  7.16  (0-136) | 1,580  7.45  (0-136) | 1,629  7.68  (0-186) | 1,518  7.16  (0-187) | **T=21,662**  **A=102.18** |
| **B** | 2,969  1.49  (0-19) | 2,372  1.19  (0-6) | 2,790  1.40  (0-8) | 2,950  1.48  (0-9) | 2,870  1.44  (0-10) | 2,910  1.46  (0-15) | 3,169  1.59  (0-16) | 3,408  1.71  (0-14) | 3,149  1.58  (0-17) | 2,651  1.33  (0-12) | 2,571  1.29  (0-11) | 3,009  1.51  (0-24) | **34,818**  **17.47** |
| **C** | 1,494  0.73  (0-21) | 1,392  0.68  (0-20) | 1,433  0.70  (0-13) | 1,494  0.73  (0-40) | 1,617  0.79  (0-49) | 1,576  0.77  (0-280) | 1,781  0.87  (0-53) | 1,617  0.79  (0-54) | 1,616  0.79  (0-34) | 1,556  0.76  (0-27) | 1,494  0.73  (0-36) | 1,638  0.80  (0-23) | **18,708**  **9.14** |
| **Sessions excluded each month across all organizations** | | **T=6,447**  **A=1.52**  **R=(0-323)** | **5,443**  **1.28**  **(0-288)** | **6,076**  **1.43**  **(0-209)** | **6,145**  **1.45**  **(0-152)** | **8,059**  **1.90**  **(0-214)** | **5,923**  **1.39**  **(0-280)** | **6,504**  **1.53**  **(0-146)** | **6,663**  **1.57**  **(0-156)** | **6,282**  **1.48**  **(0-136)** | **5,787**  **1.36**  **(0-136)** | **5,694**  **1.34**  **(0-186)** | **6,165**  **1.45**  **(0-187)** | **T=75,188**  **A=17.68** |
| **Monthly average per organization** | | **6,266 dialysis sessions with missing data per organization per month** | | | | | | | | | | | | |
| **Monthly average per facility** | | **1.47 dialysis sessions with missing data per facility per month across all organizations** | | | | | | | | | | | | |

**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*)

As a correlation between missing data and the patient-months contributed to the testing by specific patients was not conducted, we cannot identify the total percentage of dialysis sessions that were excluded over the course of the year due to missing data elements. Consequently, we instead estimate here what that proportion would be, were all patients to contribute a full year’s worth of data.

Specifically, testing involved 412,522 patients. If all patients contributed the still-conventional average of 12 to 13 dialysis sessions per month (or 144 to 156 sessions per year), the total number of dialysis sessions for all patients over the course of the year would be between 59,403,168 and 64,353,432. The total of 75,188 sessions that were excluded from the analysis due to missing data would then be 0.12 to 0.13% of all dialysis treatments.

While there is no widely held cut-off regarding an acceptable percentage of missing data in a data set for valid statistical inferences, current, conservative statistical literature suggests that a missing rate of 5% or less is inconsequential.1,2,3 Thus, even if the total patient-months were significantly less than estimated in the above calculations (i.e., secondary to deaths, hospitalizations, and transplants over the course of a year), the rate of missing data still would not be sufficient to bias performance results for the measure. For instance, if patient-months contributing to the denominator were halved from the assumptions above, the 75,188 sessions excluded due to missing data would not surpass 0.25% of all treatments for the year.

1. Dong Y and Peng CJ. Principled missing data methods for researchers. *Springerplus.* 2013;2:222-241.
2. Schafer JL. Multiple imputation: A primer. *Stat Methods in Med.* 1999;8(1):3–15.
3. Bennet DA. How can I deal with missing data in my study? *Aust N Z J Public Health*. 2001;25(5):464–469.