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

**Measure Number** (*if previously endorsed*)**:** 0251

**Measure Title**: Vascular Access—Functional Arteriovenous Fistula (AVF) or AV Graft or Evaluation for Placement

**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*).

KCQA tested its ESRD measures through a one-year prospective cohort study on a nationally drawn sample of 53 dialysis facilities and four nephrology offices.

**1.3. What are the dates of the data used in testing**? September 1, 2008-August 31, 2009

**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: 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*)

Facility Testing:

KCQA tested its ESRD measures through a one-year prospective cohort study on a nationally drawn sample of 53 dialysis facilities containing a mix of for-profit and not-for-profit providers; hospital-affiliated and freestanding facilities within large, small, and independent dialysis organizations; urban, suburban, and rural settings; and facilities both with and without electronic health records (EHRs). Facility samples were structured to be generally representative of the national industry profile as identified by the United States Renal Data Systems (USRDS) 2007 Annual Data Report.

Physician Office Testing:

To test the measure in physician offices, KCP contracted with IFMC, which was under an existing contract with the AMA PCPI/RPA to perform on-site feasibility and implementation testing of several AMA PCPI/RPA measures and had thus already obtained consent from four nephrology practice sites that would consist of a nephrology practice alpha site local to IFMC and three sites distributed geographically across the United States (Iowa, Nevada, Texas, and Pennsylvania) of various practice sizes (5.25 to 62 physicians), and medical record types (two EHR, one paper but by the time of visit transitioning to EHR, and one hybrid).

(IFMC noted that it is a Quality Improvement Organization that serves as a health oversight agency for CMS and is therefore authorized to have access to personal health information (PHI). It further noted that PHI may be disclosed to it without patient authorization under the HIPAA Privacy Rule at 45 CRF\*\*\*164.512(d).)

**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*)

Facility Testing:

Approximately 25 patients per facility were sought, resulting in a final sample size of 1,115 patients. Of these, 1,057 were hemodialysis patients and were thus included in the vascular access measures’ denominator populations. Patient samples were structured to be generally representative of the national industry profile as identified by the United States Renal Data Systems (USRDS) 2007 Annual Data Report.

Physician Office Testing:

Again, to test the measure in physician offices, KCP contracted with IFMC. Each site was asked to pull in advance the records of the first 35 adult hemodialysis patients seen on or after July 1, 2007; IFMC requested what it referred to as an oversample of five patients in an effort to ensure a remaining sample of 30 patients. The facilities within which the sample patients received care were asked to pull the records in advance of the IFMC visit because IFMC and AMA PCPI/RPA had previously identified the need for both patient’s physician office and dialysis organization records to collect necessary data elements. Physician offices were, therefore, requested to secure copies of the necessary facility records in advance of the IFMC visit.

(IFMC noted that it is a Quality Improvement Organization that serves as a health oversight agency for CMS and is therefore authorized to have access to personal health information (PHI). It further noted that PHI may be disclosed to it without patient authorization under the HIPAA Privacy Rule at 45 CRF\*\*\*164.512(d).)

**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**.

Not applicable.

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

**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*)

Facility Testing:

Following the data collection period, on-site data-integrity audits were performed at 11 of the 53 facilities (21%). Audit sites were selected to provide a cross-section of facilities reflective of the sample profile. Selection criteria included geographic location, facility type (e.g., for-profit vs. not-for-profit, urban vs. rural), and EHR use. Pertinent data were re-abstracted from the patients’ medical records and were compared to the information submitted by the facility throughout the pilot to assess the measure’s reliability.

Physician Office Testing:

The three nephrology office sites, in addition to the alpha site, were visited by a two-person IFMC abstractor team to conduct feasibility and reliability testing. Using the KCQA data collection tool, the two abstractors individually abstracted each medical record, compared the results, and evaluated the mismatches. Mismatch codes, previously developed by IFMC for reliability testing of project abstraction, were used to classify the reason determined for each mismatch.

**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*)

Facility Testing:

Inter-rater reliability was assessed during the on-site audits through a direct comparison of data submitted by the facilities throughout the pilot to data reabstracted by the auditor(s). (See Table 1 [Measure Performance, Submitted vs. Reabstracted Data] in the accompanying Attachment A.) Reliability was quantitatively summarized using Cohen’s Kappa with confidence intervals. The resulting Kappa statistic for the Functional AVF or Evaluation by Vascular Surgeon for Placement measure was found to be 0.8880 with a 95% confidence interval of 0.7484-1.000. (See Table 2 [Measure Aggregate Reliability] in Attachment A.) Based on the literature, this value indicates “almost perfect agreement” and excellent reproducibility for the measure. In addition to the Kappa value, the percent agreement between the auditor and facility abstractors (i.e., the reliability percentage) was calculated and was found to be excellent at 96.9%. (See Table 3 [Measure Reliability Percentage and Error Type] in Attachment A.) These two values demonstrate that the KCQA measure is reliable.

Physician Office Testing:

To determine whether the ESRD measure definitions and specifications, as prepared by KCQA, yield stable, consistent measurements when applied in the physician office setting, inter-rater reliability was also assessed by IFMC. As in the facility setting, the resulting Kappa statistic indicates excellent reproducibility at 0.9152 with a 95% confidence interval of 0.8349-0.9964,. (See Table 4 [Kappa Statistics with Confidence Intervals, Physician Office Setting] in Attachment A.)

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

The testing results above demonstrate excellent inter-rater agreement and high reproducibility, indicating that the measure is reliable.

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

**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)*

Validity was assessed to demonstrate that the KCQA measure specifications correctly represent and capture the data elements as intended. Following the data collection period, on-site data audits were performed at 11 of the 53 facilities (21%). Sites were selected to provide a cross-section of facilities reflective of the sample profile. Selection criteria included geographic location, facility type (e.g., for-profit vs. not-for-profit, urban vs. rural), and EHR use. Auditors compared data submitted by the facilities throughout the pilot to the information contained in the medical records so as to confirm that the submitted data were an accurate representation of the clinical records and a valid representation of what had transpired.

Additionally, KCQA posits that external validity has been met through the diligence with which the original sampling schema was crafted to reflect the national industry and patient vintage and access profiles. Because the sample is representative of the U.S. dialysis population, results can be generalized with confidence.

The KCQA measure also has both content and face validity based on the following: The measure was deemed appropriate and valid by expert opinion within the KCP and KCQA.

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

Chart Validation:

* Sensitivity = 99.38% (i.e., patients who met the measure criteria were correctly identified by the facility abstractor 99.38% of the time).
* Specificity = 85.29% (i.e., patients who did not meet the measure criteria were correctly identified by the facility abstractor 85.29% of the time).
* PPV = 96.99%% (i.e., facility abstractors accurately identified patients who met the measure criteria 96.99% of the time).
* NPV = 96.67% (i.e., facility abstractors accurately identified patients who did not meet the measure criteria 96.67% of the time).

**External Validity:**

External validity of the KCQA measures was established through the meticulous construction of patient and facility samples, modeled to reflect the national industry and patient vintage and access profiles as per the 2007 USRDS Annual Data Report of Chronic Kidney Disease & End-Stage Renal Disease, the most current volume available at the time the sample was constructed. Because the sample is representative of the U.S. dialysis population, results can be generalized with confidence.

Facility Sampling: In the United States, dialysis services are provided at more than 4,800 sites (freestanding non-profit and for-profit centers, hospital-based, and government-affiliated entities—i.e., Department of Veterans Affairs or state/county/city-run). Based on the industry profile in the 2007 U.S. Renal Data System (USRDS), a recruitment list of 71 facilities that mirrored this profile was identified so as to reach a target of 60 facilities, from which we assumed additional attrition might occur during the one-year course of data collection. Department of Veterans Affairs (VA)-affiliated and other public facilities were excluded to streamline the facility recruitment process. (VA and other public facilities represent less than two percent of dialysis sites, and less than one percent of the patient population.) Based on the USRDS data, the following target facility distribution was constructed:

• 60% from for-profit large dialysis organizations (LDO),

• 15% from non-profit LDOs,

• 20% from for-profit non-LDOs, and

• 5% from non-profit non-LDOs.

Ultimately, 53 facilities participated in the pilot. The final facility sample contained a mix of both for-profit and not-for-profit providers; hospital-affiliated and freestanding facilities within large, small, and independent dialysis organizations; urban, suburban, and rural settings; and facilities both with and without electronic medical records, and was generally representative of the national industry profile. The facility distribution in the final sample was:

• 59% from for-profit LDOs,

• 8% from non-profit LDOs,

• 21% from for-profit non-LDOs, and

• 13% from non-profit non-LDOs.

Additionally, KCP members represent approximately 85% of the community; the final sample contained facilities involved with KCP members (47 facilities; 89%) and those not (6 facilities; 11%).

Patient Sampling: Twenty-five patients per facility were sought, and three primary patient-related variables were identified: dialysis type (hemodialysis, peritoneal dialysis, or home hemodialysis), vintage on dialysis, and vascular access type. Per the 2007 USDRS report, approximately 94.5 percent of patients are on in-center hemodialysis, 5 percent on peritoneal dialysis, and 0.5 percent on home hemodialysis. The sample at the outset of the study was 92.6 percent in-center hemodialysis, 4.8 percent peritoneal dialysis, and 2.7 percent home hemodialysis. At the study’s conclusion, the profile was 92.1 percent in-center hemodialysis 5.2 percent peritoneal dialysis, and 2.7 percent home hemodialysis. (The slight overrepresentation of home hemodialysis patients resulted from the participation of a facility caring exclusively for home-based hemodialysis patients. We also note that the 2007 USRDS atlas reports on data as of the end of 2005. In fact, the home hemodialysis population has been growing, and is currently estimated by community members to be 1-2%. Thus, the actual sample more accurately reflects the current situation. Regardless, nothing in the current literature indicated this small sampling difference from the national norm would have any impact on the pilot test results, and so the pilot proceeded with the original sample rather than exclude the facility with only home hemodialysis and/or attempt to replace it.)

With respect to vintage, patients were characterized as less than 90 days, 90 days to one year, and less than one year as appropriately reflecting the relevant populations to follow the performance specified by the vascular access measures. Again, the original sample was constructed to mirror the national distribution. Based on USRDS data, this equated to 6, 11, and 8 patients per facility, respectively as of September 1, 2008.

Data collection for patients with a functional AVF (defined as using two needles in the fistula) was considerably easier than for patients without and so facilities were not permitted to self-identify patients based on AVF status. To obtain sufficient sample size to analyze the underlying purpose of the two vascular access measures, facilities were asked to select 13 patients on hemodialysis who did not have a functional AVF at the study onset.

The initial patient sample size equated to 1,325 adult patients (25 patients/53 facilities), but was reduced to 1,295 because some facilities did not have enough patients of a given type. This number was reduced to 1,115 by the study’s conclusion due to patient death, transplantation, or patient transfer out of the participating facility.

**Content and Face Validity:**

The KCQA measures have content and face validity based on the following: The measures were deemed appropriate and valid by expert opinion within the KCP and KCQA.

**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?*)

Based on the testing results, the KCQA measure has high validity. The aggregate denominator and numerator criteria have excellent sensitivity, specificity, and positive- and negative-predictive values, indicating that the measure, as specified, accurately and precisely identifies and represents the intended target populations.

The measure has strong external validity secondary to the diligence with which the original sampling schema was crafted to reflect the national industry and patient vintage and access profiles. Because the sample is representative of the U.S. dialysis population, results can be generalized with confidence.

Likewise, the measure has been deemed valid in terms of its underlying premise and construct by expert opinion within both the KCP and KCQA.

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

**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*)

**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*)

**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*)

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

**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, minimum, and maximum 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*)

The vascular access profile for the 1,057 hemodialysis patients included in the pilot sample was as follows:

• Functional AVF = 621 patients (58.8%)

• Catheter only = 306 (28.9%)

• AV graft only = 99 (9.4%)

• AVF with catheter = 26 (2.5%)

• AV graft with catheter = 5 (0.5%)

Facilities reported that 291 (86.4%) of the 337 patients who did NOT have a functional permanent access at the commencement of the study (September 1, 2008) had been evaluated by a vascular or other qualified surgeon for placement of permanent access by the conclusion of the study (August 31, 2009). Of these, 20 did not have documentation of the evaluation—a requirement to receive credit for the measure. The data elements collected thus permit calculation of performance for the measure as follows:

Performance Rate =

([Patients with AVF] + [Patients with AV graft] + [Patients without AVF or AV graft seen by surgeon for placement] – [Patients seen but without medical record documentation]) ÷ ([Total patients on hemodialysis > 90 days] – [Patients enrolled in hospice])

= (621 + 99 + 291 – 20) ÷ (1,057 – 1) = 93.8%

The performance for each individual facility in the pilot ranged from 41% to 100%, with a mean performance of 93.8%.

**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*)

As the reliability analyses indicated, the measure is specified in a manner that permits it to be reliably applied. Additionally, during the course of the pilot and during the on-site interviews, facility personnel did not report any difficulty with the measure concepts or data elements.

**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*)

Not applicable.

**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*)

Not applicable.