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

**Measure Number** (*if previously endorsed*)**:** 1667

**Measure Title**: Pediatric Kidney Disease: ESRD Patients Receiving Dialysis: Hemoglobin Level < 10g/dL

**Date of Submission**: 2/27/2015

**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*).

**1.3. What are the dates of the data used in testing**? 2007-2008

**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*)   
Four nephrology practice sites representing various types, locations and sizes were identified to participate in testing the measures

o The number of physicians per site ranged from 5-62 physicians

o The sites were located in four different regions: Midwestern, Western, Eastern, and Southern

o Patient visit volume ranged from 240-2,800 ESRD patients seen per month

• Sample size per physician organization ranged from 24-30 (as shown below) for a total of 169 ESRD patients on Peritoneal Dialysis (PD), or Hemodialysis (HD)

o Site 1: 27 ESRD patients (3 PD patients, 24 HD patients)

o Site 2 :40 ESRD patients (10 PD patients, 30 HD patients)

o Site 3 :42 ESRD patients (19 PD patients, 23 HD patients)

o Site 4 : 60 ESRD patients (30 PD patients, 30 HD patients)

• Sample selection: Data were collected from the medical records of the first up to 35 ESRD patients on each type of dialysis seen at each site after July 1, 2007.

• Data abstraction was completed for multiple patient visits per patient for a total of 2012 patient visits.

• Data abstraction was performed in 2008.

**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*)   
**Number of Records Sampled**

30 records for CKD and 30 records for ESRD were sampled at each of the 4 practice sites.

Samples of 112 CKD patients and 169 ESRD patients (62 with PD and 107 with HD), provides ample power with 5% statistical significance, power of 0.80 to 0.90, substantial agreement (kappa = 0.8) versus moderate agreement (kappa = 0.4), and trait prevalence of 0.5 to 0.75.

(See Donner and Eliasziw,1992; and Sims and Wright, 2005.)

Donner A. Eliasziw M. A goodness-of-fit approach to inference procedures for the kappa statistic: confidence interval construction, significance-testing and sample size estimation. Statistics in Medicine.1992;11(11):1511-9.

Sims J, Wright CC. "Kappa Statistics in Reliability Studies: Use, Interpretation, and Sample Size Requirements. Physical Therapy.2005;85:257-268.

**Sampling Method**

* To arrive at a sample of 30 records per condition/per site, we will over-sample for 35 records.
* For practices participating in CMS’ Physician Quality Reporting Initiative (PQRI), the sample will be drawn from a population of patients for whom: 1) a CPT II code was submitted on a claim in 2008, AND 2) the patients had at least 2 office visits with the nephrologist in the 2007 calendar year.
* For sites that are not participating in PQRI, the sample with be drawn from a patient population in which patients had at least 2 office visits with the nephrologist in the 2007 calendar year.
* IFMC will provide one of the following sampling methodologies to the practice sites:
  + Identify records for 35 patients (with CKD or ESRD) whose Social Security number ends in a specific number, i.e., 2 and 4

**OR**

* + Identify records for the first 35 patients seen during the first month of the timeframe used for testing (this method would help assure that the sampled patients would potentially have 12 months of data for review).

**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**.

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

**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*)

Data abstracted from patient records were used to calculate inter-rater reliability for the measure.

Patients were randomly selected from visits for ESRD.

Data analysis included:

• Percent agreement

• Kappa statistic with 95% confidence interval to adjust for chance agreement

http://upload.wikimedia.org/math/2/d/b/2dbc0665e308fabc49e2118c3869a6ec.png

Cohen's kappa coefficient is a statistical measure of inter-rater agreement or inter-annotator agreement for qualitative (categorical) items. It is generally thought to be a more robust measure than simple percent agreement calculation since κ takes into account the agreement occurring by chance.

**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*)  
Plan of Care for Anemia Measure (N, % Agreement, Kappa ( 95% Confidence Interval))

2012, 99.45% Agreement, Kappa 0.9860 (0.9778-0.9943)

**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?*)  
This measure is highly reliable, as shown in results from the inter-abstractor analysis (above).

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

**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)*  
An expert panel was used to assess face validity of the measure. This panel consisted of 21 members, with representation from the following specialties: nephrology, pediatric nephrology, endocrinology, nursing, methodology, internal medicine, preventive medicine and family medicine.

Louis H. Diamond, MBChB, FCP (SA), FACP, FHIMSS (Work Group Co-Chair) (Nephrology, Methodology) President, Quality Healthcare Consultants, Rockville, MD

Barbara Fivush, MD (Work Group Co-Chair) (Nephrology - Pediatrics) Professor of Pediatrics, Division Chief of Pediatric Nephrology, Johns Hopkins University, Baltimore, MD

Paul M. Palevsky, MD, FACP, FCCD, FASN (Work Group Co-Chair) (Nephrology - Adult) Professor of Medicine, University of Pittsburgh School of Medicine, Chief, Renal Section, VA Pittsburgh Healthcare System, Pittsburgh, PA

Eileen D. Brewer, MD (Nephrology - Pediatrics) Professor and Head, Pediatric Renal Section, Baylor College of Medicine Chief, Renal Service, Texas Children´s Hospital, Houston, TX

John W. Foreman, MD (Nephrology - Pediatrics) Department of Pediatrics, Professor of Pediatrics, Duke University, Durham, NC

Richard S. Goldman, MD (Nephrology - Adult, Methodology) Nephrology and Internal Medicine, Albuquerque, NM

Stuart L. Goldstein, MD (Nephrology - Pediatrics) Director, Center for Acute Care Nephrology, Cincinnati Children’s Hospital Medical Center; Medical Director, Pheresis Service, Professor of Pediatrics, University of Cincinnati College of Medicine, Cincinnati, OH

John Hartman, MD (Nephrology - Adult) CEO, Visonex, LLC, Treasurer, Wisconsin Medical Society, Green Bay, WI

Richard Hellman, MD, FACP, FACE (Endocrinology, Methodology) Clinical Professor of Medicine, University of Missouri-Kansas City School of Medicine, Private Practice, Diabetes & Endocrinology, North Kansas City, MO

Jean L. Holley, MD, FACP (Nephrology - Adult) Clinical Professor of Medicine, University of Illinois, Urban-Champaign and Carle Physician Group, Urbana, IL

Edward R. Jones, MD (Nephrology - Adult) Self-Employed, Delaware Valley Nephrology Associates, Philadelphia, PA

Karen M. Kolbusz, RN, BSN, MBA, (Nursing, Joint Commission Liaison) Associate Project Director, The Joint Commission, Oakbrook Terrace, IL

Craig B. Langman, MD (Nephrology - Pediatrics) The Isaac A. Abt MD Professor of Kidney Diseases and Head, Kidney Diseases, Feinberg School of Medicine, Northwestern University, and Children´s Memorial Hospital, Chicago, IL

Rajnish Mehrotra, MD (Nephrology - Adult) Professor of Medicine at David Geffen School of Medicine at UCLA and Associate Chief, Div of Nephrology and Hypertension, Harbor-UCLA Medical Center, Torrance, CA

Alvin H. Moss, MD (Nephrology - Adult) Professor of Medicine, West Virginia University, Morgantown, WV

Sharon A. Perlman, MD (Nephrology - Pediatrics) USF Pediatric Nephrology, All Children´s Hospital, St. Petersburg, FL

Paul D. Rockswold, MD, MPH (Preventive Medicine and Family Medicine) Physician Epidemiologist, Head of Health Analysis, Navy and Marine Corps Public Health Center, Suffolk, VA

Candace C. Walworth, MD (Nephrology - Adult) Nephrology and Internal Medicine, Lewiston, ME

Bradley Warady, MD (Nephrology - Pediatrics) Chief, Pediatric Nephrology, Children´s Mercy Hospitals and Clinics, Kansas City, MO

Steven J. Wassner, MD, FAAP (Nephrology - Pediatrics) Professor of Pediatrics, Vice-Chair for Education, Chief, Division of Nephrology & Hypertension, Hershey, PA

Jerry Yee, MD (Nephrology - Adult) Division Head, Nephrology and Hypertension, Henry Ford Hospital, Detroit, MI

Face validity of the measure score as an indicator of quality was systematically assessed as follows:

After the measure was fully specified, the expert panel (workgroup membership) was asked to rate their agreement with the following statement:

Please rate your agreement with the following statement for each measure- the scores obtained from the measure as specified will accurately differentiate quality across providers.

Scale 1-5, where 1=Strongly Disagree; 3=Neither Disagree nor Agree; 5=Strongly Agree

**2b2.3. What were the statistical results from validity testing**? (*e.g., correlation; t-test*)  
The results of the expert panel rating of the validity statement were as follows:

N =19; Mean rating = 4.37

Frequency Distribution of Ratings

1 - 1 (Strongly Disagree)

2 - 0

3 - 1 (Neither Disagree nor Agree)

4 - 6

5 - 11 (Strongly Agree)

**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?*)  
Face validity of the measure score as an indicator of quality was consistent.  
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**2b3. EXCLUSIONS ANALYSIS**

**NA**  **no exclusions — *skip to section*** [***2b4***](#section2b4)

At the time of testing, this measure did not have exclusions.

**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,** We account for risk adjustment by inclusion of the exceptions for this measure.

**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*)  
We account for risk adjustment by inclusion of the exceptions for this measure. These exceptions were added after the measure was tested and were not included in existing testing data.

**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)* Testing Project:

• Four nephrology practice sites representing various types, locations and sizes were identified to participate in testing the measures.

o The number of physicians per site ranged from 5-62 physicians.

o The sites were located in four different regions: Midwestern, Western, Eastern, and Southern.

o Patient visit volume ranged from 240-2,800 ESRD patients seen per month.

• Sample size per physician organization ranged from 24-30 (as shown below) for a total of 169 ESRD patients on Peritoneal Dialysis (PD), or Hemodialysis (HD).

o Site 1: 27 ESRD patients (3 PD patients, 24 HD patients)

o Site 2 :40 ESRD patients (10 PD patients, 30 HD patients)

o Site 3 :42 ESRD patients (19 PD patients, 23 HD patients)

o Site 4 : 60 ESRD patients (30 PD patients, 30 HD patients)

• Sample selection: Data were collected from the medical records of the first up to 35 ESRD patients on each type of dialysis seen at each site after July 1, 2007.

• Data abstraction was completed for multiple patient visits per patient for a total of 2012 patient visits.

• Data abstraction was performed in 2008.

CMS Physician Quality Reporting Initiative 2008:

“Plan of Care for ESRD Patients with Anemia” (Measure #80)

There were 179,197 eligible instances reported in the clinical performance denominator for the 2008 program. The mean clinical performance rate was 90.55% with 676 eligible professionals submitting information for the measure. The inter-quartile range (IQR) was calculated, which provides a measure of the dispersion of performance.

Data abstracted from patient records were used to calculate parallel-forms reliability for the measure.

Patients were randomly selected from visits for ESRD

Data analysis included:

• Percent agreement

• Kappa statistic to adjust for chance agreement

**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*)  
Testing Project Results:

Scores on this measure: N = 2012 Mean = 74 % Range = 61%-94%

CMS Physician Quality Reporting Initiative:

This measure was used in the CMS Physician Quality Reporting Initiative, in the claims option and the registry option (2008).

There is a gap in care as shown by this 2008 data; 36.51% of patients reported on did not receive the optimal care.

10th percentile: 10.42%

25th percentile: 38.17%

50th percentile: 66.23%

75th percentile: 84.04%

90th percentile: 94.93%

The inter-quartile range (IQR) provides a measure of the dispersion of performance. The IQR is 45.9, and indicates that 50% of physicians have performance on this measure ranging from 38.17% and 84.04%. A quarter of reporting physicians have performance on this measure which is greater than 84.04%, while a quarter have performance on this measure less than 38.17%.

Confidential CMS PQRI 2008 Performance Information by Measure. Jan-Sept TAP file.

**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?*)  
The inter-quartile range (IQR) provides a measure of the dispersion of performance. The IQR is 45.9, and indicates that 50% of physicians have performance on this measure ranging from 38.17% and 84.04%, indicating a range of performance.

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

**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*)  
 Testing Project:

• Two nephrology practice sites representing various types, locations and sizes which participated in the CMS PQRI Project in 2007 were identified to participate in testing the measures

• Sample size across the two physician offices as 74 patient visits

• Sample selection: Data were collected from the medical records of the first up to 35 ESRD patients on each type of dialysis seen at each site after July 1, 2007.

• Data abstraction was performed in 2008

Data abstracted from patient records were used to calculate parallel-forms reliability for the measure.

Patients were randomly selected from visits for ESRD

Data analysis included:

• Percent agreement

• Kappa statistic to adjust for chance agreement

**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*)  
Plan of Care for Anemia Measure (N, % Agreement)

314, 58.3% Agreement  
**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*)  
It should be noted that there were instances where the wrong quality data code was inserted on the claim based on the documented laboratory results available. This was likely due to the process of some dialysis facilities’ practice of routine billing the first of every month. This would cause a June bill to refer to laboratory results from May. This test was run in the first year of the program implementation which may have affected results as well.

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

**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*)

**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*)

**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*)