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

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

**Measure Title**: Anemia of chronic kidney disease: Dialysis facility standardized transfusion ratio (STrR)

**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*).

Data for the measure are derived from an extensive national ESRD patient database, which is derived These data are part of an extensive and comprehensive national ESRD patient database, derived from the Consolidated Renal Operations in a Web-enabled Network (CROWN) data system, Medicare claims, and the Social Security Death Master File. The CROWN data system is made up of the Renal Management Information System (REMIS) and CROWNWeb and is updated regularly using the Medicare Enrollment Database (EDB), ESRD Medical Evidence Report forms (CMS 2728), ESRD Death Notification forms (CMS 2746), and the Organ Procurement and Transplantation Network (OPTN) transplant database. The database is comprehensive for Medicare patients. Information on transfusions is obtained from Medicare Inpatient and Outpatient Claims Standard Analysis Files (SAFs).

**1.3. What are the dates of the data used in testing**? January 1, 2009 – December 31, 2012

**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*)

For each year, we first included all Medicare certified facilities. The following table shows the count of the facilities each year, before and after exclusions were applied; we also report percent excluded for each year.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Facility Count** | |  |
| **Year** | **Before Exclusions** | **After Exclusions** | **Percent Excluded** |
| 2009 | 5631 | 5478 | 2.7% |
| 2010 | 5743 | 5655 | 1.5% |
| 2011 | 5845 | 5793 | 0.9% |
| 2012 | 5936 | 5899 | 0.6% |

The following table shows the number of facilities that were included for testing and analysis for the years 2009-2012.

|  |  |  |
| --- | --- | --- |
| **Year** | **# Facilities** | **Mean Facility size (patients)** |
| 2009 | 5478 | 68.99 |
| 2010 | 5655 | 69.88 |
| 2011 | 5793 | 69.93 |
| 2012 | 5899 | 71.97 |

**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*)

|  |  |  |  |
| --- | --- | --- | --- |
| **Year** | **# Facilities** | **# of Patients** | **Total Patient Years at Risk** |
| 2009 | 5478 | 377957 | 215069.1 |
| 2010 | 5655 | 395183 | 223699.0 |
| 2011 | 5793 | 405104 | 227598.8 |
| 2012 | 5899 | 424563 | 233794.9 |

The following table shows the facility level mean number of patients, mean age; mean values for patient years at risk, mean %females , %black, %white, and %Hispanics for each of the four years.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Facility Level Mean values** | | | | | | |
| **Year** | **# Patients** | **Age as of end of year** | **Patient Yrs at Risk** | **%Female** | **%Black** | **%White** | **%Hisp** |
| 2009 | 68.99 | 61.53 | 39.01 | 45.76 | 36.24 | 56.90 | 14.63 |
| 2010 | 69.88 | 61.66 | 39.41 | 45.72 | 36.25 | 56.93 | 14.69 |
| 2011 | 69.93 | 61.72 | 39.10 | 45.52 | 36.02 | 57.13 | 14.90 |
| 2012 | 71.97 | 61.69 | 39.44 | 45.34 | 36.03 | 57.01 | 14.84 |

**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**.

All reliability, validity, risk adjustment analyses are done using this data set as explained in Table 2 of Section 1.5 above.

For testing of meaningful differences (please refer section 2b.5), as implemented in the NQF endorsed Standardized Hospitalization Ratio measure (NQF #1463 http://www.qualityforum.org/QPS/1463) facilities with less than 10 patient years at risk are excluded from this analysis.

The table below gives the counts of facilities included or excluded facilities, with less than 10 patient years at risk for the years 2009-2012.

|  |  |  |  |
| --- | --- | --- | --- |
| **Year** | **# Facilities Included in the Testing and Analysis** | **# Facilities with at least 10 patient years at risk** | **Percent excluded** |
| 2009 | 5478 | 4798 | 12.4% |
| 2010 | 5655 | 4986 | 11.8% |
| 2011 | 5793 | 5118 | 11.7% |
| 2012 | 5899 | 5279 | 10.5% |

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

**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*)

The reliability of the STrR was assessed using data among ESRD dialysis patients during 2009-2012. If the measure were a simple average across individuals in the facility, the usual approach for determining measure reliability would be a one-way analysis of variance (ANOVA), in which the between and within facility variation in the measure is determined.The inter-unit reliability (IUR) measures the proportion of the measure variability that is attributable to the between-facility variance. The STrR, however, is not a simple average and we instead estimate the IUR using a bootstrap approach, which uses a resampling scheme to estimate the within facility variation that cannot be directly estimated by ANOVA. A small IUR (near 0) reveals that most of the variation of the measures between facilities is driven by random noise, indicating the measure would not be a good characterization of the differences among facilities, whereas a large IUR (near 1) indicates that most of the variation between facilities is due to the real difference between facilities.

Here we describe our approach to calculating IUR. Let *T*1,…,*TN* be the STrR for these facilities. Within each facility, select at random and with replacement *B* = 100 bootstrap samples. That is, if the *i*th facility has *ni* subjects, randomly draw with replacement *ni* subjects from those in the same facility, find their corresponding STrR*i* and repeat the process 100 times. Thus, for the *i*th facility, we have bootstrapped STrRs of ,…, Let be the sample variance of this bootstrap sample. From this it can be seen that

is a bootstrap estimate of the within-facility variance in the STrR, namely, . Calling on formulas from the one way analysis of variance, an estimate of the overall variance of *T*i is

where

is the weighted mean of the observed STrR and

is approximately the average facility size (number of patients per facility). Note that is an estimate of , where is the between-facility variance, the true signal reflecting the differences across facilities. Thus, the IUR, which is defined by

can be estimated with

**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 STrR calculation only included facilities with at least 10 patient years at risk. Overall, we found that IURs for the one-year STrR have a range of 0.49-0.55 across the years 2009, 2010, 2011 and 2012, which indicates that around half of the variation in the one-year STrR can be attributed to the between-facility differences and half to within-facility variation. This value of IUR indicates a **moderate degree of reliability**. When stratified by facility size, we find that, as expected, larger facilities have greater IUR.

**Table: IUR for One-year STrR, Overall and by Facility Size, 2009-2012**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 2009 | | 2010 | | 2011 | | 2012 | |
| Facility Size  (No. of patients) | IUR | N | IUR | N | IUR | N | IUR | N |
| All | 0.49 | 4797 | 0.53 | 4985 | 0.55 | 5117 | 0.54 | 5278 |
| Small (<=46) | 0.36 | 1513 | 0.44 | 1576 | 0.38 | 1706 | 0.36 | 1743 |
| Medium (47–78) | 0.46 | 1637 | 0.49 | 1682 | 0.52 | 1687 | 0.54 | 1817 |
| Large (>=79) | 0.59 | 1647 | 0.6 | 1727 | 0.66 | 1724 | 0.65 | 1718 |

**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 temporal positive correlation values of the measure indicates that facilities with higher (or lower) transfusion rates in one year tend to have higher (or lower) transfusion rates in the following year.

This value of IUR indicates a moderate degree of reliability. When stratified by facility size, we find that, as expected, larger facilities have greater IUR.

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

**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)*

Validation of critical data elements: The critical data elements for this measure (transfusion and exclusion comorbidity claims) come from Medicare claims data. However we did not conduct validation for each individual data element as it is an exhaustive process.

Validation of the measure: We assessed validity of the STrR by performing Spearman correlations between the STrR and other measures of quality among ESRD population. Results of the correlation estimates were statistically significant. In May 2012 there was an assessment of the measure’s face validity based on polling of a CMS Technical Expert Panel (TEP).

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

The validity of the STrR measure is supported by its association with other known quality measures, which include both dialysis facility outcomes and practices. Spearman’s rho is reported for all measures. For year 2012, we find that the measure is positively correlated with two health outcome measures: the NQF endorsed #1463 one-year Standardized Hospitalization Ratio for Admissions (rho = 0.40, p < .0001), the NQF endorsed #0369 one-year Standardized Mortality Ratio (rho = 0.23, p < .0001), and the one-year Standardized Readmission Ratio (rho = 0.17, p < .0001). We also tested the association with average hemoglobin values of all ESA-treated dialysis patients (rho = -0.16, p < .0001). The negative correlation indicates that lower values of hemoglobin are associated with higher values of STrR. Similarly, the positive correlation with the percent of patients with Hgb < 10 (rho = 0.20, p < .0001) indicates that a higher % of patients with Hgb < 10 is associated with higher STrR.

Furthermore, the STrR is positively correlated with catheter use (rho =0.22, p < .0001), indicating that higher values of STrR are associated with increased use of catheters. The STrR is negatively correlated with the percentage of patients with Kt/V>=1.2 (rho = -0.09, p < .0001) and using a fistula (rho = -0.08, p < .0001). That is, higher values of STrR are associated with lower rates of achievement of adequate dialysis measured by Kt/V>=1.2 and fistula use.

Six out of six voting members of CMS’s 2012 Technical Expert Panel voted to recommend development of a facility-level Standardized Transfusion Ratio measure. The consensus recommendation of that clinical expert panel included the recommendation to include risk adjustment for conditions that are associated with an increased risk of blood transfusion such as hereditary anemia, chronic bone marrow failure conditions and active cancer.

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

The overall measure demonstrates both strong face validity and construct validity. The positive correlation between this measure and SMR and SHR respectively indicates that facilities with more transfusions than would be expected based on national rates, also have higher standardized mortality and standardized hospitalization rates, i.e., standardized to the national rates.

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

**2b3. EXCLUSIONS ANALYSIS**

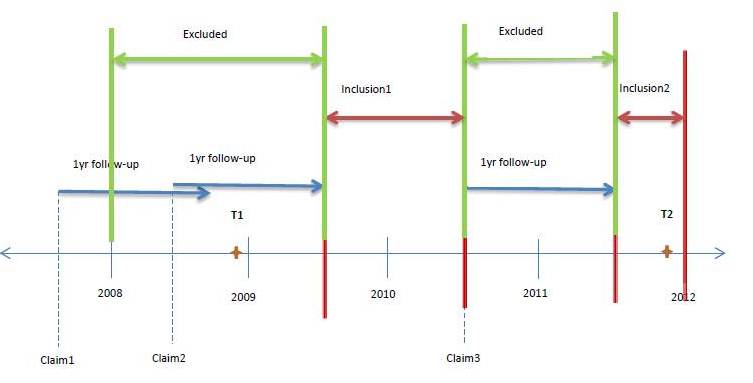
**NA**  **no exclusions — *skip to section*** [***2b4***](#section2b4)

**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*)

Transfusions associated with transplant hospitalization are excluded as they mark a transition of care from the dialysis facility to a transplant team. This convention is used with other dialysis facility measures developed and previously endorsed by NQF (like SHR NQF #1463 http://www.qualityforum.org/QPS/1463) and SMR NQF #0369 http://www.qualityforum.org/QPS/0369)

Patients are also excluded if they have a Medicare claim for hemolytic and aplastic anemia, solid organ cancer (breast, prostate, lung, digestive tract and others), lymphoma, carcinoma in situ, coagulation disorders, multiple myeloma, myelodysplastic syndrome and myelofibrosis, leukemia, head and neck cancer, other cancers (connective tissue, skin, and others),metastatic cancer, sickle cell anemia within one year of their patient at risk time. Since these comorbidities are associated with higher risk of transfusion and require different anemia management practices that this measure is not intended to address, every patient’s risk window is modified to have at least 1 year free of claims that contain diagnoses on the exclusion list. We assessed the predictive power of comorbidities on future transfusions, as a function of the time interval between development of the comorbidity and the occurrence of the transfusion by performing multivariate logistic regression with transfusion event as the dependent variable. Results showed that 1-year look back period for each of the above mentioned comorbidities was the most predictive of one or more RBC transfusions.

The following figure describes the inclusion and exclusion period of a hypothetical patient.



In the figure above, a hypothetical patient has patient years at risk at a facility from 1/1/2008 to 12/31/2011. Review of Medicare claims identified presence of one or more exclusion comorbidities (see above and Appendix)e in 2007 (Claim1), 2008 (Claim2) and 2010 (Claim3). Each claim is followed by a one year exclusion period. The revised inclusion periods are defined as risk windows with at least 1 year of claim-free period (Inclusion1 and Inclusion2 in the figure). The patient has two transfusion events, marked as T1 and T2 in late 2008 and late 2011 respectively. However, since T1 falls in the exclusion period, it will not be counted towards the facility’s transfusion count as presence of exclusion comorbidity claims within a year might have increased the risk of transfusion unrelated to dialysis facility anemia management practice. However, T2, which occurs in late 2011 and in Inclusion2 period, will be counted since there is at least a year gap between this transfusion event and the last claim observed.

**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*)

Multivariate logistic regression with transfusion count as the dependent variable was performed to assess the predictive power of comorbidities on future transfusions, as a function of the time interval between development of the comorbidity and the occurrence of the transfusion. Transfusion event was coded as a binary variable (1 if transfusion). Result using 2011 data showed that 1-year look back period for each of the above mentioned comorbidities was a significant predictor of RBC transfusion events with odds ratio ranging from 1.2 to 3.2.

The following tables show percent of patient years at risk and number of patients excluded as a result of the above mentioned exclusion strategy.

Table 1: Percent of patient years at risk excluded each year

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Patient years at risk** | |  |
| **Year** | **Before Exclusion** | **After Exclusion** | **Percent** |
| 2009 | 267086 | 215069.1 | 19.5% |
| 2010 | 278073 | 223699.0 | 19.6% |
| 2011 | 286821 | 227598.8 | 20.6% |
| 2012 | 295172 | 233794.9 | 20.8% |

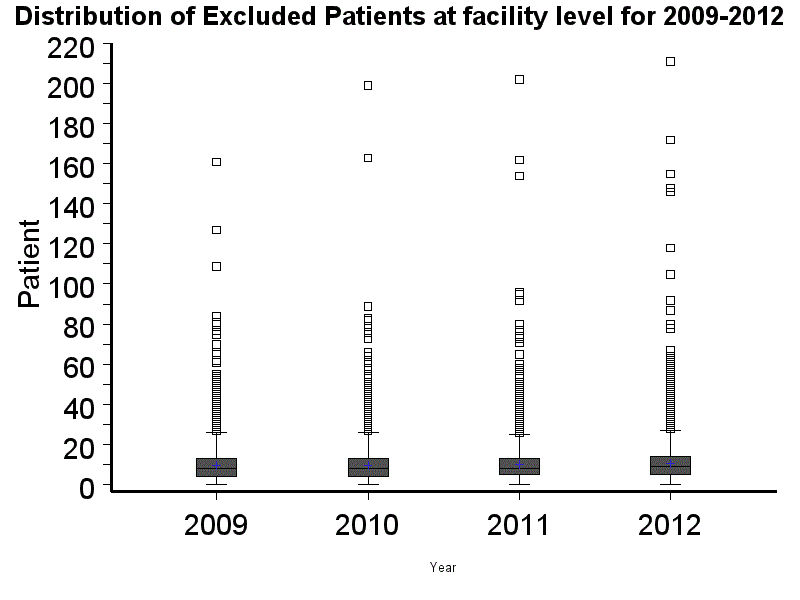
Table2: Number of patients and percent excluded each year

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Number of Patients** | |  |
| **Year** | **Before Exclusion** | **After Exclusion** | **Percent** |
| 2009 | 438949 | 377957 | 13.9% |
| 2010 | 459395 | 395183 | 14.0% |
| 2011 | 475044 | 405104 | 14.7% |
| 2012 | 502056 | 424563 | 15.4% |

**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*)

The list of comorbidities described in section 2b3.1 are associated with higher risk of transfusion and require different anemia management practices that this measure is not intended to address; hence the need for the comorbidity exclusions. The Technical Expert Panel had also recommended these exclusions. As described in Section 2b3.2 patients with exclusion comorbidities are at a higher risk to get transfused than others.

We also checked the distribution of patients excluded at the facility level and the boxplot shows that there is variability in the number of patients excluded among facilities. The numbers of patients with the exclusion comorbidities are not uniformly distributed across facilities thereby demonstrating the need for an exclusion strategy.



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

**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** 7 **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**.

N/A

**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 included all the standard patient demographics that are included in the facility level modeling for primary outcomes. We sought input from clinicians and epidemiologists and incorporated claims based risk factors and covariate adjustments recommended by the Technical Expert Panel. We also ensured that the variables are chosen to harmonize with those that are used as risk-adjusters for a similarly constructed measure, Standardized Hospitalization Ratio, which is NQF endorsed (http://www.dialysisdata.org; NQF #1463 http://www.qualityforum.org/QPS/1463)

The denominator of the “STrR” is an estimate of the expected number of transfusions at the facility; accounting for each patient’s follow-up time and risk factors. The expected number of transfusions is based on the recurrent event analog of Cox regression (Cox, 1972), as developed by Lawless and Nadeau,(1995) and Lin et al. (2000); see also Kalbfleisch and Prentice, 2002). For computational purposes, we adopt a model with piecewise constant baseline rates (e.g. Cook and Lawless, 2007) and computational methodology as developed in Liu, Schaubel and Kalbfleisch (2010).

The calculation of the STrR is a two-stage approach. At Stage 1, model is first fitted to the national data with piecewise-constant baseline rates stratified by facility; transfusion rates are adjusted for patient age, diabetes, duration of ESRD, nursing home status, BMI at incidence, comorbidity index at incidence, and calendar year. This model allows the baseline transfusion rates to vary between strata (facilities), but assumes that the regression coefficients are the same across all strata; this approach is robust to possible differences between facilities in the patient mix being treated. The regression parameter estimates from Stage 1 are used to compute the expected number of transfusions for each patient. Stage two involves summing the expected number of transfusions by facility, then computing facility-specific STrRs as the ratio of observed / expected transfusions.

The patient characteristics included in the stage 1 model as covariates are

* age (18-24 years old, 25-44 years old, 45-59 years old, 60-74 years old, or 75+ years old)
* cause of ESRD (diabetes or other)
* nursing home status
* BMI at incidence
* comorbidity index at incidence
* duration of ESRD (91 days-6 months, 6 months-1 year, 1-2 years, 2-3 years, 3-5 years, or 5+ years as of the period start date) and
* calendar year

Nursing home status is identified as in or not in a nursing home in the previous calendar year. The comorbidity index is calculated as a weighted linear combination of comorbidities reported on the Medical Evidence Form (CMS-2728) namely alcohol dependence, atherosclerotic heart disease, cerebrovascular disease, chronic obstructive pulmonary disease, congestive heart failure, diabetes, diabetes (currently on insulin), drug dependence, inability to ambulate, inability to transfer, malignant neoplasm, cancer, other cardiac disease, peripheral vascular disease, tobacco use (current smoker) using the same weights as used for Standardized Hospitalization Ratio (http://www.dialysisdata.org; NQF #1463 http://www.qualityforum.org/QPS/1463). BMI is included as a log-linear term. Categorical indicator variables are included as covariates in the stage 1 model to flag records with missing values for cause of ESRD, comorbidity index, and BMI. These variables have a value of 1 if the patient is missing the corresponding piece of information and a value of 0 otherwise. Another categorical indicator variable included as a covariate to flag records where the comorbidity index is 0 has a value of 1 if the patient has a comorbidity index of 0 and a value of 0 otherwise.

Beside main effects, some two way interaction terms are also included in the model based on their clinical and statistical significance.

**2b4.4. What were the statistical results of the analyses used to select risk factors?**In the table below, we list results from the Stage 1 model described above. For a given covariate, the parameter estimate represents the log of the rate ratio (recurrent event version of the relative risk). All covariates have face validity from a clinical perspective; none describe quality of care; and none are associated with disparities in care. With the exception of Cause of ESRD=missing (included for the purposes of completeness), all main effects are statistically significant at 0.05 level.

| Covariate | Coefficient | P-value |
| --- | --- | --- |
| Incident comorbidity index |  |  |
| 0 | -0.127 | <.0001 |
| Incident comorbidity index (continuous) | 0.375 | <.0001 |
| Missing | -0.068 | 0.012 |
| Cause of ESRD |  |  |
| Diabetes | -0.075 | <.0001 |
| Missing | -0.038 | 0.063 |
| Age |  |  |
| 18-24 | 0.0187 | 0.312 |
| 25-44 | -0.234 | <.0001 |
| 45-59 | -0.169 | <.0001 |
| 60-74 | Reference |  |
| 75+ | 0.008 | 0.213 |
| BMI |  |  |
| Log BMI | -0.193 | <.0001 |
| BMI missing | 0.108 | <.0001 |
| Calendar year |  |  |
| 2009 | Reference |  |
| 2010 | -0.033 | <.0001 |
| 2011 | -0.040 | <.0001 |
| 2012 | -0.067 | <.0001 |
| In nursing home the previous year | 0.542 | <.0001 |
| Diabetes as cause of ESRD & time on ESRD interaction term |  |  |
| 91 days-6 months | Reference |  |
| 6 months-1 year | 0.072 | <.0001 |
| 1-2 years | 0.102 | <.0001 |
| 2-3 years | 0.127 | <.0001 |
| 3-5 years | 0.078 | <.0001 |
| 5+ years | 0.058 | <.0001 |
| Age & diabetes as cause of ESRD interaction term |  |  |
| 0-14 |  |  |
| 15-24 | 0.26 | 0.002 |
| 25-44 | 0.272 | <.0001 |
| 45-59 | 0.126 | <.0001 |
| 60-74 | Reference |  |
| 75+ | 0.012 | 0.176 |
|  | | |

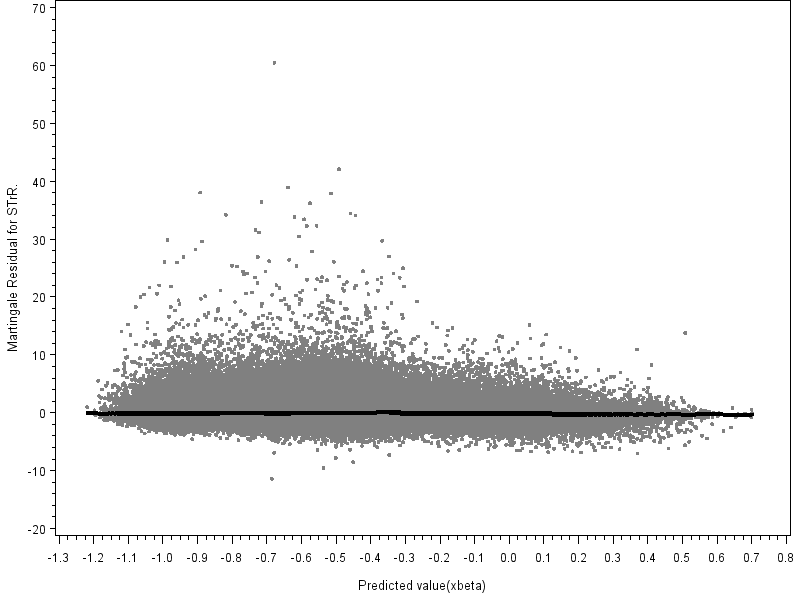
**2b4.5. Describe the method of testing/analysis used to develop and validate the adequacy of the statistical model or stratification approach** (*describe the steps―do not just name a method; what statistical analysis was used*)

*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)

Martingale residuals (Barlow and Prentice, 1988) are an important tool for checking the fit of a Cox regression model (or, a model analogous to a Cox model; including the one we fitted at Stage 1). Martingale residual plots are used to investigate the lack of fit of a model. We examined the residual plots and it did not indicate problems with the model fit. The LOESS curve of martingale residuals by predicted value (Figure 1) shows that the mean of the residuals is flat indicating no lack of fit.

Reference: Barlow, W. E. and Prentice, R. L. (1988). Residuals for relative risk regression. Biometrika 75, 65{74.

Figure 1: Martingale Residual for STrR



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

We looked into goodness of fit statistics, specifically at AIC, BIC criterion. For the model with all the covariates as specified above, AIC value is 3774215.2, as compared to the model with intercept only for which AIC = 3798267.8. Smaller AIC is better. Similar results are obtained when comparing BIC values for both the models. The AIC and BIC values reflect great importance of the adjustment covariates, in aggregate.

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

We ranked each subject based on their average expected event rate. We then broke the subjects up into deciles and computed decile-specific observed and expected numbers of transfusions. Results are given in the table below; with the relative agreement between the observed and expected counts given in the last column. Overall, the model appears to have good calibration.

|  |  |  |  |
| --- | --- | --- | --- |
| **Decile** | **Observed transfusions** | **Expected transfusions** | **(Obs- Exp)/Exp** |
| 1 | 29709 | 30485.12 | -0.025 |
| 2 | 34139 | 34604.26 | -0.013 |
| 3 | 33935 | 34499.89 | -0.016 |
| 4 | 36107 | 36321.78 | -0.006 |
| 5 | 37053 | 37453.20 | -0.011 |
| 6 | 37716 | 38150.72 | -0.011 |
| 7 | 39199 | 39176.86 | 0.001 |
| 8 | 40689 | 40395.72 | 0.007 |
| 9 | 43904 | 42313.23 | 0.038 |
| 10 | 62057 | 61516.19 | 0.009 |

**2b4.8. Statistical Risk Model Calibration – Risk decile plots or calibration curves**:  
Decile plots (Figure2) shows piecewise linear estimates of the cumulative rates by years since start of ESRD. The plot demonstrates that the risk factors in the model are discriminating well between patients. There is good separation among all 10 groups and the ordering is as predicted by the model (patients predicted to be at lower risk have lower transfusion rates). The absolute differences between the groups is also large with patients predicted to have the highest transfusion rates (line 10) having almost 3 times higher transfusion rates than those predicted to have the lowest rates (line 1).

Figure 2: Decile plots for count of transfusions

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

N/A

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

Covariates used as risk adjusters for this measure all have face and clinical validity and most of them are statistically significant at the 0.05 level. The residual plots show no lack of fit, while goodness-of-fit criteria show that there is great value in risk adjustment. The model appears to adequately discriminate the risk of transfusion among subjects; and, overall, is well-calibrated.

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

**N/A**

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

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

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

The STrR is a ratio of the observed number of red blood cell transfusions to the expected number among patients in a facility over a 1‐year . The expectation is obtained based on the overall national average rate of transfusions, adjusted for the particular patient mix at the facility under consideration.

In order to classify facilities as having transfusion rates that are better, no different or worse than the national average, we require a method of obtaining a p‐value for classification purposes. A p-value assesses the probability that the facility would experience a number of transfusions more extreme than that observed if the null hypothesis were true; accounting for each facility’s patient mix. To do this, a z‐score is first calculated using the estimate and standard error for each facility using the method of generalized estimating equations (GEE; Liang & Zeger, 1986). Specifically, the transfusion rate (or, equivalently: the mean transfusion count, given the exposure) was assumed to follow a multiplicative model and a robust (sandwich) standard error was used. The use of robust standard errors has been advocated for modeling recurrent events (i.e., multiple events per subject) by several previous authors; e.g., Lawless & Nadeau (1995); Lin, Wei, Yang & Ying (2000); Cai & Schaubel (2004). For each facility, the Z-score was computed as the facility’s log(STrR), divided by its standard error. Since log(STrR) is undefined for facilities with 0 transfusions, the Z-score in such cases was computed as (STrR-1), divided by a standard error estimate (sandwich estimator) for STrR.

To account for the over dispersion in the z-scores, as used in Standardized Hospitalization Ratio (NQF #1463 http://www.qualityforum.org/QPS/1463), we use robust estimates of location and scale based on the center of the z-scores (by fitting robust regression on z- scores) and derive normal curves that more closely describes the z‐score distribution. This new distribution is referred to as the “empirical null hypothesis” (Efron, 2004) and provide references for assessing the extent to which a given facility’s outcomes are extreme in comparison with other facilities. We then use the mean and standard deviation from the empirical null distribution of the STrR z‐scores to calculate the p‐value for classifying facility performance**.**

References:

* Lin, D.Y., Wei, L.J., Yang, I. and Ying, Z. (2000). Semiparametric regression for the mean and rate functions of recurrent events. Journal of the Royal Statistical Society Series B, 62, 711–730.
* Cai, J. and Schaubel, D.E.. (2004). Marginal means and rates models for multiple-type recurrent event data. Lifetime Data Analysis, 10, 121-138.
* Liang, K.Y. and Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models. Biometrika, 73, 13-22.
* Lawless, J.F. and Nadeau, C. (1995). Some simple robust methods for the analysis of recurrent events. Technometrics, 37, 158-168.
* Efron, B. (2004). Large scale simultaneous hypothesis testing: the choice of null hypothesis. J. Amer. Statist. Assoc., 99, 96‐104.

**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 following Table shows how the facilities are flagged for the year 2012, based on the method described above.

**Table 1: Classification of Empirical p-value for year 2012**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Year 2012** | **Frequency** | **Percent** | **Cumulative Frequency** | **Cumulative Percent** |
| **Better than expected** | 23 | 0.44 | 23 | 0.44 |
| **As expected** | 4907 | 92.95 | 4930 | 93.39 |
| **Worse than Expected** | 349 | 4.05 | 5279 | 100 |

**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 results tabulated above indicate that the STrR has the ability to classify facilities as being significantly better (or significantly worse) than expected; thereby demonstrating the ability to identify meaningful differences in the performance scores across facilities.

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

**2b6. COMPARABILITY OF PERFORMANCE SCORES WHEN MORE THAN ONE SET OF SPECIFICATIONS**

***If only one set of specifications, this section can be skipped.***

**Note***: This criterion is directed to measures with more than one set of specifications/instructions (e.g., one set of specifications for how to identify and compute the measure from medical record abstraction and a different set of specifications for claims or eMeasures). It does not apply to measures that use more than one source of data in one set of specifications/instructions (e.g., claims data to identify the denominator and medical record abstraction for the numerator).* ***If comparability is not demonstrated, the different specifications should be submitted as separate measures.***

**2b6.1. Describe the method of testing conducted to demonstrate comparability of performance scores for the same entities across the different data sources/specifications** (*describe the steps―do not just name a method; what statistical analysis was used*)

N/A

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

N/A

**2b6.3. What is your interpretation of the results in terms of demonstrating comparability of performance measure scores for the same entities across the different data sources/specifications?** (i*.e., what do the results mean and what are the norms for the test conducted*)  
N/A

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

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

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

N/A

**2b7.2. What is the overall frequency of missing data, the distribution of missing data across providers, and the results from testing related to missing data?** (*e.g.,**results of sensitivity analysis of the effect of various rules for missing data/nonresponse; if no empirical sensitivity analysis, identify the approaches for handling missing data that were considered and pros and cons of each*)  
N/A

**2b7.3. What is your interpretation of the results in terms of demonstrating that performance results are not biased** due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias**?** (i*.e., what do the results mean in terms of supporting the selected approach for missing data and what are the norms for the test conducted; if no empirical analysis, provide rationale for the selected approach for missing data*)

N/A