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

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

**Measure Title**: Optimal End Stage Renal Disease (ESRD) Starts

**Date of Submission**: Click here to enter a date

**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: Data request forms faxed back by dialysis facilities and kidney transplant centers |

**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*).

The Data Submission Form (appendix) is completed by each of the participating Kaiser Permanente (KP) regions. The information is abstracted from care coordination databases of ESRD patients/members. Since initiation of dialysis and transplantation outside of KP require authorization, it is believed that no patients are missed. The Data Submission Forms are combined into a single database for calculation of Optimal ESRD Starts and the results are reported as total Optimal ESRD Starts and its components (home dialysis, preemptive kidney transplant, in-center hemodialysis via fistula, graft and catheter).

For validity testing of the data elements, a randomized sample set from the above dataset was selected (as described in 1.6). For each randomized patient, a data questionnaire was mailed to the dialysis facility where the patient initiated dialysis or to the kidney transplant program where the patient was transplanted. The questionnaire requested verification that the patient had indeed started dialysis or received a kidney transplant and the date of initiation of dialysis or transplantation (verifying denominator inclusion), if dialysis, the dialysis modality and if hemodialysis, the vascular access used for the first treatment (catheter, arteriovenous fistula or arteriovenous graft). Forms that were not returned by deadline were followed up by phone calls and faxes to the facility. All but two were completed and returned; those facilities were no longer under KP contract and claimed they no longer had the information. This information was tested against the information on the Data Submission Form (above).

**1.3. What are the dates of the data used in testing**? Calendar year 2012; 01Jan-31Dec 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: KP Regions: Integrated delivery system | **☒** other: KP Regions: integrated delivery system |

**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*)

All six (6) KP regions participating in the Optimal ESRD Program are included in the validity testing and analysis. Regional membership distribution, in thousands, is 226, 236, 482, 538, 3,385, and 3,566. Regions are located in northern California, southern California, Hawaii, northern Oregon and southern Washington, Colorado, and Georgia.

**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*)

All patients who met the optimal ESRD starts criteria during the analysis period are included in the sampling pool.

A stratified random sample proportionally allocated across the strata (regions) based upon the number of qualifying (see measure specifications section) patients in the region for the measurement period, calendar year 2012, is selected based upon an element match probability of .9 with a .95 confidence level on a bound of .1 (lower bound of .8 and upper bound of 1).

The number of records is doubled from 35 to 70 to compensate for potentially missing data and rounded up when allocating to the regions for a total of 73 records.

Sample count distribution across regions is 3, 3, 3, 4, 28, and 32.

**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**.

Only one sample is drawn and used for data element validity testing.

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

**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.*

Accuracy/correctness (validity) of data elements is empirically tested. This section, 2a2, is not required and is skipped.

**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*)

**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*)

**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?*)

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

**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)*

We test the accuracy of the regional data compared to the authoritative source comparing the renal replacement therapy information submitted by the regional care coordinator (region) to that provided by the renal replacement therapy provider on record (source), i.e. dialysis unit where the initial dialysis occurred (hemodialysis) or managed (peritoneal dialysis) or the transplant center on record as performing the kidney transplant.

Among the data element elements, we test our a priori hypothesis of a .9 match (90% accuracy) between the region and source data. This is performed at three levels, denominator, numerator, and total.

1. At the denominator level, we test the accuracy of the qualification elements, for example, a first time dialysis. We measure the proportion of records where the region response and source matched and use a 95% confidence interval (CI) to determine if the estimated proportion is different from our hypothesized value of .9.
2. At the numerator level, we compare the method of renal replacement submitted by the region to the authoritative source. This is to assess the accuracy of the regional data element identifying the mode of initial renal replacement therapy. A 95% CI is used to determine if the estimated proportion differs from our hypothesized value of .9.
3. Total element accuracy, denominator and numerator accuracy combined, is tested comparing the match proportion to our hypothesized value of .9 and a 95% CI.
4. We evaluate the performance metric, optimal ESRD start, using sensitivity, specificity, and positive and negative predictive values with 95% confidence intervals.
5. There are two sample cases where the source, dialysis center, is no longer under contract. For these two cases the source results are not available. To judge the potential effect of the missing data, we consider two scenarios, (1) the region result does not match the source (worst case), and (2) the region result does match the source (best case).

Data element match proportion confidence intervals are calculated such that the lower confidence limit is never less than zero and the upper confidence limit is never greater than one (Fleiss).

Fleiss JL (1981). Statistical Methods for Rates and Proportions, 2nd edition. John Wiley & Sons, New York.

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

Among the 73 sample records, there are 12 errors comprised of 9 errors at the numerator level, where the region recorded mode of renal replacement does not match the source record (AV Fistula and Catheter), and 3 errors at the denominator level, where the patient does not meet the qualifying criteria (does not start renal replacement or has previously started treatment). Two records have missing source results due to the dialysis center not currently under contract, Table 1.

For the denominator, numerator, and total element match proportions, Table 2, we exclude the two records with no source results (scenario 1, exclude). To consider the potential effect of the excluded records, we include the records under two additional scenarios, that the region result does not match the source (scenario 2, as error) and where the region result matches the source (scenario 3, as match).

1. For the denominator match, the match proportion is .96 (68/71). This is the proportion of sample records where the initial regional submission matches the result from re-examination. The 95% lower and upper confidence limits are .87 and .99, respectively. Under the alternative scenarios, the denominator accuracy of the region is .93 (68/73) with CI .84-.98, if the missing validation data are considered errors and .96 (70/73) with CI .88-.99 if the two missing results are considered matches.
2. For the numerator match, the match proportion is .87 (59/68). The three denominator record errors are excluded. The 95% lower and upper confidence limits are .76 and .93. Assuming that the two cases without source data do not match the source, the region accuracy is .84 (59/70) with CI .73-.92. Assuming that the regional data match the source, then the accuracy is .87 (61/70) with CI .76-.94.
3. For the total element match, the match proportion is .83 (59/71). The 95% lower and upper confidence limits are .71 and .91. Assuming that the region results does not match the missing source results, the match proportion is .81 (59/73) and CI .70-.89. Assuming that the regional data matches the source, then the accuracy is .84 (61/73) with CI .73-.91.
4. Collapsing Table 1 using the optima ESRD start definition yields the distribution for the performance metric, Table 3, a 2 x 2 table. The region results are displayed in the rows and source in the columns. For example, among the total of 68 cases, the regions list 35 of the cases as having an optimal ESRD start. Of these 35, 33 are true optimal ESRD start cases as reported by the source. The region result matches the source in 59 cases, 33 where region and source are both ‘Yes’ and 26 where region and source are both ‘No’. Table 3 is used to calculate the optimal ESRD starts test results, Table 4.
5. For each scenario, Table 4 lists the regional accuracy (match proportion), prevalence of a true (source) optimal ESRD start in the sample, sensitivity and specificity of the regional data to correctly identify an optimal ESRD start among true optimal ESRD start cases and to correctly identify cases that are not an optimal ESRD start among such cases, positive and negative likelihood ratios, and the positive and negative predictive values. For the observed data, excluding the missing data, the match proportion is .87 (59/68) with CI 0.76-0.93, prevalence of a true (source) optimal ESRD start is .59 (40/68). Sensitivity and specificity are .82 with CI (0.67, 0.92) and .93 with CI (0.75-0.99) respectively. Positive and negative predictive values are .94 with CI (0.80, 0.99) and .79 with CI (0.61, 0.90).
6. Results for the missing data scenarios are discussed in section 2b7 Missing Data Analysis and Minimizing Bias.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 1. Distribution of treatment counts for region and source. | | | | | | | | |
|  | Source (Dialysis Unit/Transplant Center) | | | | | Not first dialysis | No source result | Total |
| Region | AV Fistula | Catheter | Peritoneal | Transplant | Graft |
| AV Fistula | 19 | 2 |  |  |  | 1 |  | 22 |
| Catheter | 7 | 26 |  |  |  | 2 | 2 | 37 |
| Peritoneal |  |  | 11 |  |  |  |  | 11 |
| Transplant |  |  |  | 3 |  |  |  | 3 |
| Graft |  |  |  |  | 0 |  |  | 0 |
| Total | 26 | 28 | 11 | 3 | 0 |  |  | 73 |
| Notes | Not first dialysis: Error for Optimal ESRD Start qualifications, exclude from denominator. No source result: Source information not available, exclude from analysis. | | | | | | | |
|

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Table 2. Match proportion results | | |  |  |  |  |
|  |  | Denominator | Numerator | Proportion | 95% Confidence Interval | |
| Match | Scenario | n | x | p = x/n | LCL | UCL |
| Total | Exclude | 71 | 59 | 0.831 | 0.719 | 0.906 |
|  | As error | 73 | 59 | 0.808 | 0.696 | 0.888 |
|  | As match | 73 | 61 | 0.836 | 0.727 | 0.909 |
| Numerator | Exclude | 68 | 59 | 0.868 | 0.759 | 0.934 |
|  | As error | 70 | 59 | 0.843 | 0.732 | 0.915 |
|  | As match | 70 | 61 | 0.871 | 0.765 | 0.936 |
| Denominator | Exclude | 71 | 68 | 0.958 | 0.873 | 0.989 |
|  | As error | 73 | 68 | 0.932 | 0.841 | 0.975 |
|  | As match | 73 | 70 | 0.959 | 0.877 | 0.989 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Table 3. Distribution of Optimal ESRD Start counts for Region and Source. | | | | |
|  | Source | |  |  |
| Region | Yes | No | Total |  |
| Yes | 33 | 2 | 35 |  |
| No | 7 | 26 | 33 |  |
| Total | 40 | 28 | 68 |  |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 4. Optimal ESRD Start metric test results: match proportion, prevalence, sensitivity, specificity, predictive value (95% confidence interval) | | | | | | | | |
|  |  |  |  |  |  |  | Predictive Value | |
| Scenario | Proportion | Prevalence | Sensitivity | Specificity | LR+ | LR- | Positive | Negative |
| Exclude | 0.868 (59/68) | 0.588 | 0.825 | 0.929 | 11.5 | 0.2 | 0.943 | 0.788 |
|  | (0.759, 0.934) |  | (0.666, 0.921) | (0.75, 0.988) |  |  | (0.795, 0.99) | (0.606, 0.904) |
| As error | 0.843 (59/70) | 0.571 | 0.786 | 0.867 | 5.9 | 0.3 | 0.892 | 0.743 |
|  | (0.732, 0.915) |  | (0.628, 0.892) | (0.684, 0.956) |  |  | (0.736, 0.965) | (0.564, 0.869) |
| As match | 0.871 (61/70) | 0.600 | 0.833 | 0.933 | 12.5 | 0.2 | 0.946 | 0.800 |
|  | (0.765, 0.936) |  | (0.68, 0.925) | (0.765, 0.988) |  |  | (0.805, 0.991) | (0.625, 0.909) |

**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?*)

At the element level, the accuracy of region data compared to the source is good to excellent with a match in the denominator of .96, in the numerator of .87, and total match of .83. At the 95% confidence level these proportions are not significantly different from our hypothesized proportion of .9. The region data is a valid representation of the source at the element level.

At the performance metric level, accuracy is very good where the region assessment matches the source with a proportion of .87 (59/68) and the 95% CI encompasses values from good to excellent (.76, .93). With a sensitivity proportion of .82, the region assessment is very good at identifying cases that are true optimal ESRD starts with a 95% confidence that the accuracy is within .67 to .92. Specificity, the probability that the region assessment correctly identifies cases that are not optimal ESRD starts, is excellent, .93.

The probability of a true optimal ESRD start among cases identified by the region as optimal ESRD start, the positive predictive value (PPV), is excellent at .94. Conversely, identification of a true non-optimal ESRD start among region reported cases as non-optimal ESRD starts, negative predictive value (NPV), is good at .79. Both PPV and NPV are dependent upon prevalence of true optimal ESRD starts. As the use of optimal ESRD starts increase (prevalence), the PPV increases and NPV decreases.

A third pair of test assessments, based upon the sensitivity and specificity values, are the likelihood ratios. This pair is independent of the prevalence of the optimal ESRD start. The positive likelihood ratio (LR+), sensitivity/(1 – specificity), is the odds of a true optimal ESRD start among region cases identified as an optimal ESRD start (sensitivity) compared to region cases identified as a non-optimal ESRD start ((1-specificity) is 11.6. That is, the region correctly identifies a true optimal ESRD start 11.6 times more often than it incorrectly identifies a non-optimal ESRD start as optimal. This is more than double the generally accepted rule that a test should have a ratio greater than 5. The negative likelihood ratio (LR-), calculated as (1-sensitivity)/specificity, is the odds of a true optimal ESRD start among cases identified by the region as non-optimal compared to region correctly identified non-optimal ESRD starts among non-optimal. The resulting odds, .19, meets the generally accepted threshold that a LR- should be less than 0.2.

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

**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*)  
NA. No exclusions.

**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*)  
NA. No exclusions.

**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*)  
NA. No exclusions.

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

**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)***.***

The optimal ESRD start is a process measure. This section, 2b4, is skipped per instructions.

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

☐ **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**. NA. This is a process measure.

**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*) NA

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

**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*) NA

*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*)**:** NA

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

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

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

**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*) NA

**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*) NA

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

**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).*

To identify statistically significant and meaningful differences in the optimal ESRD start metric, we compare the regions to the all regions (regional aggregate) performance using CY2012 study performance results. The validity sample is drawn from the same CY2012 population. The regional optimal ESRD start data, number of cases and proportion receiving an optimal ESRD start, is used to test for significant differences among and between the regional performance and the all regions performance. A chi-square test is performed to test if there are any regional differences from the all regions performance. A second set of tests to determine which region is different compares each regional optimal ESRD start proportion to the all regions proportion and a 95% confidence interval is used to determine statistical significance.

**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*)

For the chi-square test, the calculated statistic is 29.73 with 5 degrees of freedom. This is statistically significant (p-value is < .0001).

Results comparing the region to all region optimal ESRD start proportion are in Table 6. The table includes the number of cases, the number of optimal ESRD starts, optimal ESRD start proportion, difference from the all regions performance and the 95% confidence interval. One region, Region 1, is identified as significantly different from the all regions performance. For region 1, among the 110 cases newly diagnosed as ESRD in CY2012, 31 received an optimal ESRD start as identified by the region for a start proportion of .28, difference from the all regions is -.22 with a 95% CI (-0.32, -0.13).

Table 6. Optimal ESRD Start results by region, measurement period CY 2012.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Region** | **N** | **Optimal ESRD Start Count** | **Opt ESRD Start Proportion** | **Difference** | **LCL** | **UCL** |
| 1 | 110 | 31 | 0.282 | -0.224 | -0.315 | -0.133 |
| 2 | 87 | 42 | 0.483 | -0.023 | -0.136 | 0.090 |
| 3 | 76 | 32 | 0.421 | -0.085 | -0.204 | 0.035 |
| 4 | 1031 | 516 | 0.500 | -0.005 | -0.042 | 0.031 |
| 5 | 134 | 76 | 0.567 | 0.061 | -0.029 | 0.151 |
| 6 | 1162 | 618 | 0.532 | 0.026 | -0.009 | 0.061 |
| All Regions | 2600 | 1315 | 0.506 |  |  |  |

Table 7. Optimal ESRD Start, region to national (.355\*) performance comparison, measurement CY2012 using 95% confidence intervals.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Region** | **N** | **Optimal ESRD Start Count** | **Opt ESRD Start Proportion** | **LCL** | **UCL** | **Significance+** |
| 1 | 110 | 31 | 0.282 | 0.202 | 0.377 | NS |
| 2 | 87 | 42 | 0.483 | 0.375 | 0.592 | Sig |
| 3 | 76 | 32 | 0.421 | 0.310 | 0.540 | NS |
| 4 | 1031 | 516 | 0.500 | 0.470 | 0.531 | Sig |
| 5 | 134 | 76 | 0.567 | 0.479 | 0.652 | Sig |
| 6 | 1162 | 618 | 0.532 | 0.503 | 0.561 | Sig |
| All Regions | 2600 | 1315 | 0.506 | 0.486 | 0.525 | Sig |

\* Estimate based upon USRDS and CMS Fistula First data, see Appendix for calculation

+ NS, Not significantly different, 95% confidence interval contains the national value, LCL < national value < UCL; Sig, Significantly different, 95% confidence interval does not include national value.

**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 in Tables 6 and 7 indicate that there is sufficient variation in regional results to identify significant differences in performance among regions both within Kaiser Permanente and to the national optimal ESRD start rate.

The Table 6 results indicate that the optimal ESRD start proportions vary by region (.282-.567) and the regional difference to the all regions group mean difference can be statistically significant (Region 1, difference = -.224, significant using 95% confidence level). Thus indicating that identification of performance differences is possible with the optimal ESRD start metric.

The individual region and the aggregate, all region, results are compared to the 2012 national rate (0.355, estimated from the USRDS and CMS Fistula First data, see Appendix for the calculation), Table 7, using the 95% confidence intervals for the region and all region optimal ESRD start rates. Four of the regional (2, 4, 5, and 6) and the all region aggregate confidence intervals do not include the national optimal ESRD start performance value, .355 indicating statistically significant differences from the national rate.

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

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

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

**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*) NA

**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*) NA

**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*) NA

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

**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*)

To quantify the effect of missing data, the performance metric is calculated under the two scenarios (1) for the excluded cases, the region data elements and optimal ESRD start result are assumed not to match the source; and (2) where the region data elements and optimal ESRD start result are assumed to match the source. These two scenarios represent the two extremes for the missing data.

Missing data are only amongst the numerator data elements, the method of renal replacement.

**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*)

There are two cases with missing source data. Both cases are the result of the dialysis center not being under contract during the validation period with one case from each of the two largest regions. One missing case is from a national dialysis chain center and the other from an independent dialysis center. No dialysis center has more than one record in the study sample. Sample selection uses a random selection of cases across the entire CY2012 population of ESRD patients newly diagnosed in the period. The proportion of missing cases (2/73) is less than 3%.

Sensitivity analysis results for the data elements are displayed in Table 2. For the total element match, the observed match proportion can range from a low of .81 when the region data elements do not match the source for the missing cases (Scenario = As error) to a high of .84 when the region data elements match the source (Scenario = As match). When the region does not match the source, the match proportion is significantly different from .9 at the 95% level (CI, .70-.89). For the numerator match, the match proportion has a low of .84 and a high of .87. For the denominator, the match proportion is .93 when the region does not match the source and .96 when it does. The numerator and denominator extremes are not statistically different from the a priori match proportion, .09.

Performance metric results are in Table 4. The optimal ESRD start metric is .84 (95% CI .73-.92) under the ‘As error’, both sensitivity and specificity drop to .78 and .87 respectively, the positive likelihood ratio is now 6 with a negative LR increasing to .3. For the ‘As match’ scenario, the performance metric is .87 (95% CI .77-.94), sensitivity is .83 (95% CI .68-.93), specificity is .93 (.77-.99), positive LR increases slightly to 12.5 and negative LR remains at .2.

**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*)

Given the relatively small number of cases with missing data, less than 3% (2/73), we expect a small effect to be seen in the sensitivity analysis and because the match proportion is greater than .5 that any differences are greater when missing data are treated as errors than when treated as matches. This is seen the Table 2 and 4 results. The largest change is seen in the positive LR where the observed ratio is 12 compared to 6 in the ‘As error’ scenario. For all of the other test metrics, the differences are much smaller. Using the simple comparison of confidence intervals, all of the ‘As error’ and ‘As match’ confidence intervals overlap within the respective columns indicating that there is no statistical difference between the ‘As error’ and ‘As match’ value. The missing data has no statistically significant effect upon the observed results.