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

**Measure Number** (*if previously endorsed*)**:** 0256

**Measure Title**: Minimizing Use of Catheters as Chronic Dialysis Access

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

**Type of Measure:** Intermediate outcome

|  |  |
| --- | --- |
| 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*).

We analyzed national CROWNWeb data from January 2013-December 2013.

For section 2b6, we compare 2013 CROWNWeb data to 2013 Medicare claims data.

**1.3. What are the dates of the data used in testing**? January 2013-December 2013

**1.4. What levels of analysis** **were tested**? (*testing must be provided for all the levels specified and intended for measure implementation, e.g., individual clinician, hospital, health plan*)

|  |  |
| --- | --- |
| **Measure Specified to Measure Performance of:**  **(*must be consistent with levels entered in item S.26*)** | **Measure Tested at Level of:** |
| individual clinician | individual clinician |
| group/practice | group/practice |
| hospital/facility/agency | hospital/facility/agency |
| health plan | health plan |
| other: 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*)

Patients on maintenance hemodialysis during the last HD treatment of month including patients on home dialysis from January 2013- December 2013 were included in the analyses. The number of facilities ranged from 5,592-5,670 and the total number of patient-months per month ranged from 348,199-358,263.

Public reporting of this measure on DFC or in the ESRD QIP would be restricted to facilities with at least 11 eligible patients for the measure. We have applied this restriction to all the reliability and validity testing reported here.

**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*)

There were a total of 4,226,245 eligible patient-months. Among those patient-months over the whole year, the average age was 63 years, 44% of patient-months were female, 55% were white, 38% were black, 4% were Asian, 17% were Hispanic, and 46% had type II diabetes as the primary cause of ESRD.

**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**.

N/A

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

**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*)

We used January 2013 – December 2013 CROWNWeb data to calculate facility level monthly and annual performance scores. We assessed reliability by calculating inter-unit reliability (IUR) for each reporting month and the overall 12 months. The monthly based measure was a simple average across individuals in the facility. The NQF-recommended approach for determining measure reliability is 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 yearly based measure, 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. We note that the method for calculating the IUR was developed for measures that are approximately normally distributed across facilities. Since this measure is not normally distributed, the IUR value should be interpreted with some caution.

We also did a comparison of the data elements used to calculate this measure in order to assess comparability of the calculations using two Medicare data sources. The measure was calculated using Medicare claims and CROWNWeb clinical data. The method of this comparison is further described in 2b6.1.

**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*)

For reliability we calculated the monthly and annual IUR across the 12 reporting months. As explained above, the method for calculating the IUR was developed for measures that are approximately normally distributed across facilities.  IUR=0.77813 which is high and suggests 78% of variation in the measure is attributed to between facility variation.

The results of our comparison between Medicare claims and CROWNWeb can be found in 2b6.2.

**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 IUR suggests this measure is reliable. However, since the distribution of performance scores is skewed, the IUR value should be interpreted with some caution.

The interpretation of the results of our comparison between Medicare claims and CROWNWeb can be found in 2b6.3.

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

**2b2. VALIDITY TESTING**

**2b2.1. What level of validity testing was conducted**? (*may be one or both levels*)  
 **Critical data elements** (*data element validity must address ALL critical data elements*)

**Performance measure score**

**Empirical validity testing** **Systematic assessment of face validity of performance measure score as an indicator** of quality or resource use (*i.e., is an accurate reflection of performance on quality or resource use and can distinguish good from poor performance*)

**2b2.2. For each level of testing checked above, describe the method of validity testing and what it tests** (*describe the steps―do not just name a method; what was tested, e.g., accuracy of data elements compared to authoritative source, relationship to another measure as expected; what statistical analysis was used)*

Validity was assessed using Poisson regression models to measure the association between facility level quintiles of performance scores and the 2013 Standardized Mortality Ratio (SMR, NQF 0369) and 2013 Standardized Hospitalization Ratio (SHR, NQF 1463), respectively, both NQF endorsed measures. Facility-level performance scores were divided into quintiles and the relative risk (RR) of mortality (and hospitalization, separately) was calculated for each quintile. The fifth quintile was used as the reference group. Thus, a RR>1.0 for the lower performance score quintiles would indicate a higher relative risk of mortality or hospitalization.

**2b2.3. What were the statistical results from validity testing**? (*e.g., correlation; t-test*)  
Quintiles of the performance scores were defined as follows:

Q1: 0.0%-<4.2%

Q2: 4.2%-<7.0%

Q3: 7.0-<9.7%

Q4: 9.7%-<13.6%

Q5: 13.6%-<60.6%

Results from the Poisson model indicated that the percent of patient-months dialyzing with a catheter >= 90 days was significantly associated with both SMR (p<0.0001) and SHR (p<0.0001). For 2013 SMR, relative risk of mortality was the lowest in the lowest performance measure quintile (RR=0.89; 95% CI: 0.86, 0.91). For quintile 2, RR=0.92 (95% CI: 0.90, 0.94), quintile 3, RR=0.92 (95% CI: 0.90, 0.94) and was 0.94 for quintile 4 (95% CI: 0.92, 0.96).

Similarly for 2013 SHR, the relative risk of hospitalization increased as the performance measure quintile increased from quintile 1. For quintile 1, RR=0.82 (95% CI: 0.81, 0.82), quintile 2, RR=0.89 (95% CI: 0.89, 0.90), quintile 3, RR=0.92 (95% CI: 0.91, 0.92) and was 0.95 for quintile 4 (95% CI: 0.94, 0.95).

**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?*)

These results of the Poisson regression suggest the predictive relationship of lower catheter use with lower mortality and hospitalization, as measured by the respective standardized mortality and hospitalization rates, compared to facilities with higher catheter use (>=90 days).

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

**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*)

N/A

**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*)

N/A

**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*)

N/A

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

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

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

**No risk adjustment or stratification**

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

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

**Other,** Click here to enter description

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

Trend analysis suggests disparities in sex, race, ethnicity and age, however the conservative interpretation would be that these differences reflected in the trend analysis reflect disparities in care for certain subpopulations. In the absence of biological effects explaining these differences, risk adjustment for these factors would potentially mask disparities in care.

**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*)

N/A

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

N/A

**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*)

N/A

*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*)**:**

N/A

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

N/A

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

N/A

**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*)

N/A

**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)*

Differences in measure performance were evaluated separately for each facility using patient level analyses. For each facility, the proportion of patient-months with Catheter >= 90days, calculated at the year-level, was compared to the overall national distribution.

We use a re-sampling-based exact test, with re-sampling generated from the population distribution of the patient level outcomes. More details for the testing method are provided in Appendix. Due to the non-symmetric structure of the measure distributions, a one-sided test with significance level 0.025 is used (corresponding to a cutoff=0.05 in a two-sided test). To calculate the p-value, we assess the probability that patients in each facility would experience a number of events (i.e., months dialyzing with Catheter >= 90 days) more extreme than what was actually observed if the null hypothesis were true, where the null hypothesis is that a patient in each facility will follow the overall national distribution.

**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*)

Proportion of facilities with significant differences (p-values < 0.025) is shown as follows:

Category Number of facilities Percent of facilities

As expected 5086 88.3%

Worse than expected 677 11.7%

**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?*)

Using monthly percentage of patients with Catheter more than 90 days, calculated at the year-level, as the performance measure, 5086 (88%) facilities have achieved expected performance, and 677 facilities (12%) have performed worse than expected (higher catheter prevalence).

In general, higher rates of Catheter use more than 90 days represents worse quality of care. This analysis demonstrates both practical and statistically significant differences in performance across facilities based on their proportion of patient months with Catheter more than 90 days.

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

**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*)

The measure was calculated for the time period of January – December 2013. Monthly calculations for this measure were calculated from CROWNWeb monthly clinical extracts at a patient- and facility- level. The numerator and denominator definitions were coded to match as closely as possible to the current definitions in the DFC Medicare Claims codes. In some cases, definitions were amended to account for differences in data elements in the new data source. Vascular access-type categories differed slightly between Medicare Claims and CROWNWeb. For instance, CROWNWeb categories included all the Claims categories but, had an extra category for ‘Port’ and a category for ‘Fistula & Graft’. We classified ‘Fistula & Graft’ patients as “Multiple access” and classified ‘Port’ patients as “Catheter”. Calculation of at least 90 days with a catheter also differed.

Monthly CROWNWeb calculations were summarized to quarterly and yearly time periods and merged with the DFC Medicare Claims files. For the purposes of comparing the measures by data source, all patients in these two files were divided into three mutually exclusive groups. Patients found in both files were defined as “CROWNWeb and Claims” patients. Patients found only in CROWNWeb files were considered “CROWNWeb Only”. Similarly, patients unique to Medicare Claims were classified as “Claims Only”. The CROWNWeb Only population is not currently included in the DFC measures, but would be included if CROWNWeb became the underlying data source. The opposite is true for the Claims Only population; it is currently included in the DFC measures, but would not be included if CROWNWeb became the data source.

To assess agreement between the two data sources at the patient-month-level with respect to the performance score, the number and percentage of patient-months with Catheter >= 90 days in CROWNWeb were compared to the corresponding number and percentage from Medicare Claims on a year-level (Table 1). The Kappa statistic is displayed as a footnote to each table. Table 2 displays the results at the facility level. The Pearson product-moment correlations are displayed in the table.

**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*)

**Table 1. Number and Percent of HD Patient-months in Catheter >= 90 Days Category in CROWNWeb and Claims for the Year (January - December 2013)**

|  |  |  |  |
| --- | --- | --- | --- |
| **CROWNWeb** | **Claims** | | |
| **Catheter >= 90 days** | | **Total** |
| **Catheter >= 90 days** | **YES** | **No** |  |
| **Yes** | 286482 (9%) | 61779 (2%) | 348261 (11%) |
| **No** | 69569 (2%) | 2804315 (87%) | 2873884 (89%) |
| **Total** | 356051 (11%) | 2866094 (89%) | 3222145 (100%) |

\*Percentages in parentheses are of all patient-months in the analysis.

Kappa for data agreement among all patient-months in the testing period is 0.79 (p <0.001).

Table 1 reports rates of patient-month-level catheter greater than or equal to 90 days both at 11%. Agreement is strong as demonstrated by a Kappa of 0.79 (p<0.001).

**Table 2. Facility Level Measure: Mean Percent of Catheter >= 90 Days in CROWNWeb and Medicare Claims by Year and Quarter (January - December 2013)**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **N of Facilities** | **Catheter >= 90 Days** | | **Correlation (ρ)** |
| **CROWNWeb** | **CLAIMS** |
| **Year** |  |  |  |  |
| 01/13-12/13 | 5,735 | 11.9 | 11.0 | 0.78\* |
| **Quarter** |  |  |  |  |
| 01/13-03/13 | 5,613 | 12.1 | 11.5 | 0.70\* |
| 04/13-06/13 | 5,618 | 11.9 | 11.3 | 0.70\* |
| 07/13-09/13 | 5,618 | 11.8 | 10.6 | 0.77\* |
| 10/13-12/13 | 5,669 | 11.7 | 10.3 | 0.75\* |

\*p<0.001

At the facility level, using patients in both CROWNWeb and Medicare Claims, the mean percentage of patients in the facility with Catheter>= 90 days are reported in Table 2. The mean percentages across each quarter for both data sources are comparable and the correlation coefficients for the catheter performance score calculated with CROWNWeb and claims are strong and range from 0.70 to 0.77.All are statistically significant.

**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*)

In Table 1, we see strong agreement in catheter scores between CROWNWeb and Medicare Claims as the data sources, as demonstrated by the Kappa of 0.79 (p <0.001). This is for the entire testing period (January - December 2013). Similarly, in Table 2 we observe strong correlations between the measures calculated with each source, in the one-year period and by quarter; all statistically significant. Both sets of results suggest additional evidence of reliability for the Catheter >= 90 days measure.

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

**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