**National Quality Forum—Measure Testing (subcriteria 2a2, 2b2-2b6)**

**Measure Title**: OAF-02 Risk Assessment/Treatment After Fracture

**Date of Submission**: 12/6/2013

**Type of Measure:**

|  |  |
| --- | --- |
| ☐ Composite – ***STOP – use composite testing form*** | ☐ Outcome (*including PRO-PM*) |
| ☐ Cost/resource | X☐ 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.  **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.    **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**.  **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:** |
| X☐ abstracted from paper record | X☐ abstracted from paper record |
| ☐ administrative claims | ☐ administrative claims |
| ☐ clinical database/registry | ☐ clinical database/registry |
| X☐ abstracted from electronic health record | X☐ abstracted from electronic health record |
| ☐ eMeasure (HQMF) implemented in EHRs | ☐ eMeasure (HQMF) implemented in EHRs |
| ☐ other: Click here to describe | ☐ other: Click here to describe |

**1.2. If an existing dataset was used, identify the specific dataset** (*the dataset used for testing must be consistent with the measure specifications for target population and healthcare entities being measured; e.g., Medicare Part A claims, Medicaid claims, other commercial insurance, nursing home MDS, home health OASIS, clinical registry*).

**1.3. What are the dates of the data used in testing**? January and February 2007, August through October 2009, and November 1, 2011-April 30, 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 |
| X☐ hospital/facility/agency | X☐ 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*)   
52 United States acute care hospitals participated In the alpha test for validity in 2009.

Geographically, the majority of respondents were from the Northeast and North Central segments of the country; there were virtually no volunteers from the Great Plains states, and the Pacific Northwest and Southeast segments of the country were each represented by a handful of volunteers. There were 6 volunteers from California. 6 site visits were conducted at a sample of these hospitals chosen to reflect a variety of geographic distribution, bed size and other characteristics and were located in the states of California, Connecticut, Illinois, Maryland, Ohio, and North Carolina. During these site visits, roundtable discussions were held with a variety of practitioners who cared for osteoporosis and fracture patients.

Individuals who evaluated the measures were from the following disciplines:

* Quality Improvement 38
* Physician 37
* Abstraction/Data Staff 34
* Administration 32
* Nursing Manager 21
* Pharmacist 13
* Nursing Staff 12
* Education 10
* Other 19

**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*)   
In the pilot test for reliability, records were submitted for 2049 patients. All patients were over the age of 49; males and females were included. Racial distribution is not known because the data were de-identified but all races were included. Diagnoses were those specified for each measure in Tables 5.1 and 6.1 in the attached Excel file.

**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**.

Validity testing was conducted in an alpha test that preceded the reliability test. Validity was assessed by focus group sessions at a sample of United States hospitals that had previously completed an online questionnaire; there were no patient records examined in the validity test. Reliability testing was subsequently conducted on 2049 de-identified patient records.

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

**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*)  
X☐ **Critical data elements used in the measure** (*e.g., inter-abstractor reliability; data element reliability must address ALL critical data elements*)   
X☐ **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*)

Twenty three hospitals from fifteen states volunteered to participate in a six month pilot test of this measure, commencing with discharges beginning November 1, 2011 and concluding on April 30, 2012. Twelve of the twenty three hospitals withdrew prior to the start of testing and during the first month of testing, citing lack of resources to complete the project. Pilot test hospitals ranged in size from 99 to 600 beds. Most of the eleven remaining test hospitals used some combination of paper and electronic records; most other hospitals used fully electronic records that were not integrated but contained in multiple legacy systems.

The objectives of the pilot test were:

* Assessment of data element and measure reliability
* Assessment of data collection and implementation effort
* Identification of potential measure specification enhancements

A web-based data collection tool developed by The Joint Commission was used. Joint Commission staff conducted educational and training webinars for participating test sites. The Technical Advisory Panel Chairperson also presented information to participants in one of the educational sessions. Technical staff assisted hospitals with data collection issues, ensuring that each data submission was complete, comprising good quality data.

Monthly conference calls with the pilot sites were conducted. Questions from participating hospitals could also be posted on the web-based tool where answers were posted within 24 hours. Test participants also shared tools for implementing policy and procedure changes that supported the test measures, and these tools were also posted on the web-based site.

In order to achieve the test objectives, 3 evaluations were conducted during the 6 month test.

1. Reliability Test

A reliability test was carried out at a subset of the eleven participating hospitals. Six hospitals selected to represent geographic, bed size, type, and ownership diversity, were visited by teams of two Joint Commission staff during February, March, April and May 2012. During these visits, Joint Commission staff re-abstracted records previously abstracted by hospital staff. A total of 133 records were re-abstracted. Reliability was addressed at the data element level by calculating an agreement rate between the original abstraction and the re-abstraction. In cases of disagreement between the re-abstracted and original data elements, differences were adjudicated and the reasons for disagreements were coded. Data element agreement rates for re-abstracted and adjudicated rates were calculated.

1. Data Collection and Implementation Effort

A questionnaire was distributed to each test hospital, encouraging assessment of time spent in data collection and time spent in implementing any of the measures.

1. Validity Test

A separate questionnaire was distributed to each hospital to obtain feedback regarding any enhancements to the measure specifications, or to the measures themselves, that would contribute to ease and accuracy of data abstraction and clinical approaches. These items were also discussed during pilot test site visits.

These methods assess the completeness and adequacy of measure and data element definitions, accuracy and completeness of data abstraction instructions, and accuracy of coding and other specifications.

**2a2.3. For each level 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*)  
Reliability testing was performed at the data element and measure level at 6 hospitals on a total of 133 records. The reliability (Kappa) score for Measure 02 was 0.835 (excellent); the overall match rate was 91%. The match rates for measure-specific data elements are shown below:

|  |  |  |
| --- | --- | --- |
| **Data Element** | **No. Mismatches** | **Match Rate** |
| DXA Scan Ordered or Performed Prior to Discharge | 2 | 98.50% |
| Other Fracture Risk Assessment Method Ordered or Performed Prior to Discharge | 2 | 98.50% |
| FDA-approved Pharmacotherapy for Treatment of Osteoporosis | 2 | 98.50% |
| Reason for No DXA Scan | 3 | 97.74% |
| Reason for No FDA-approved Pharmacotherapy for Treatment of Osteoporosis | 3 | 97.74% |

**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?*)  
Measure 02 shows excellent reliability (above 0.75).

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

**2b2. VALIDITY TESTING**

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

☐ **Performance measure score**

X☐ **Empirical validity testing**X☐ **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 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)*  
To determine feasibility and identify areas for potential revision, test sites were asked to electronically rate the clarity of numerator statements, denominator statements, and measure information forms (MIFs) on a five point Likert scale (1 = very poor, 2 = poor, 3 = average, 4 = good, 5 = very good). Data elements and associated tables were evaluated for clarity, collectability, and applicability of data sources.

In addition, focus group discussions were held at all test sites visited, during which we received feedback as to whether the measure, data elements, and definitions accurately reflected existing evidence. All of the respondents indicated that all aspects of the measures accurately reflected current evidence.

Please note with respect to ICD-10-CM/PCS codes:

1. Goal is to convert this measure to a new code set, fully consistent with the intent of the original measure.
2. Please refer to the Excel spreadsheet attached at data field **S.2b. Data Dictionary or Code Table**
3. The ICD-9-CM to ICD-10-CM/PCS translation process was managed by a coding expert, Mariruth Petrik, MBA, RHIA, CCS, CCS-P. The process of identification of the ICD 10 codes was done as follows. The codes were identified using the 3M GEM’s FY 2013 tool, running a forward map. These codes were then reviewed for clinical intent by the Registered Nurse responsible for the Osteoporosis measure set, as well as the Registered Health Information Administrator. Extraneous codes were eliminated to preserve the intent of the measure.

**2b2.3. What were the statistical results from validity testing**? (*e.g., correlation; t-test*)  
Average ratings of MIFs (measure information forms), numerator and denominator statements are shown in Table 2. The highest rating was 4.33 and the lowest was 2.83. Ratings are displayed in Table 2, below.

Table 2 Measure Validity Ratings

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Measure Name | MIF Under-  standing | Numerator Clarity | Numerator Inclusion Clarity | Numerator Exclusion Clarity | Denom. Clarity | Denom. Inclusion Clarity | Denom. Exclusion Clarity |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| OAF-02 Assess/  Treat After Fracture | 2.83 | 3.97 | 3.18 | 4.33 | 2.97 | 3.23 | 3.56 |

Evaluation of data elements by the method described in2b2.2 is displayed in Table 3.

Table 3 Data Element Validity Ratings

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Date Element Name** | **Applicable To** | **Clear** | **Collectable** | **Data Sources Applicable** |
| Admission Date | All measures | 100% | 100% | 97.9 % |
| Birthdate | All measures | 100% | 100% | 93.5% |
| Clinical Trial | All measures | 89.4% | 72.3% | 68.1% |
| Comfort Measures Only | All measures | 93.6% | 87.2% | 89.4% |
| Discharge Date | All measures | 100% | 100% | 95.7% |
| ICD9-CM Other Diagnosis | All measures | 100% | 100% | 97.8% |
| ICD9-CM Principal Diagnosis | All measures | 100% | 100% | 93.3% |
| Bone Mineral Density Test Performed in the 12 Months Prior to the Fracture | 02 | 91.3% | 54.3% | 60.9% |
| FDA-approved Pharmacotherapy for Prevention or Treatment of Osteoporosis | 02 | 93.2% | 93.3% | 86.4% |
| Other Fracture Risk Assessment Method Ordered or Performed Prior to Discharge | 02 | 82.2% | 77.3% | 77.3% |
| Reason for No DXA Scan | 02 | 87% | 80.4% | 77.3% |
| Reason for No FDA-approved Pharmacotherapy for Prevention or Treatment of Osteoporosis | 02 | 87% | 82.6% | 77.3% |

**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?*)  
We note that ratings for both data elements and measure are relatively high; most numerators, denominators and measures were ranked above the midpoint of 3.0 (average), and data elements were above 75% positive in clarity, collectability, and correctness of data sources. We conclude that the measures and specifications are valid.

**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*)  
 A web-based tool was created by The Joint Commission. Hospitals participating in the pilot test abstracted information from each patient record and entered the appropriate allowable values for each data element into this tool. Data elements included those that were exclusions to the measure population. At six participating hospitals, 30 records were chosen at random for re-abstraction by Joint Commission staff, who reviewed the same records and re-abstracted data into the tool. After re-abstraction, data entered by the hospital was compared to data entered by Joint Commission staff. Discrepancies were identified and resolved

**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*)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Measure | Exclusion | Overall Occurrence N | Overall Occurrence % | Distribution Across Hospitals 10th, 50th, 90th percentile |
| Osteo-02 | Discharged Home From The ED | 581 | 9.5% | 0.0, 0.06, 0.15 |
| Osteo-02 | Age | 1 | 0.0% |  |
| Osteo-02 | Other Diagnosis | 161 | 2.6% | 0.0, 0.02, 0.10 |
| Osteo-02 | Discharge Disposition | 23 | 0.4% | 0.0, 0.0, 0.01 |
| Osteo-02 | Comfort Measures Only | 33 | 0.5% | 0.0, 0.00, 0.01 |
| Osteo-02 | Clinical Trial | 19 | 0.3% | 0.0, 0.0, 0.002 |
| Osteo-02 | Bone Mineral Density Test Performed in the 12 Months Prior to the Fracture | 18 | 0.3% | 0.0, 0.00, 0.01 |
| Osteo-02 | On FDA-approved Pharmacotherapy for Treatment of Osteoporosis Prior to Fracture | 96 | 1.6% | 0.0, 0.01, 0.04 |
|  |  |  |  |  |

**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*)  
Except for patients Discharged from the ED (9.5%) and those with an Other Diagnosis (2.6%), the frequency of exclusions is low, but we believe all exclusions should be retained for the following reasons:

Discharged Home from the ED and Discharge Disposition– Measure is intended for inpatients; care prescribed is not possible to give in an ED without significantly raising costs an increasing ED throughput time. Additionally, care for detecting causes of fracture or treating a chronic illness is not within the scope of practice of an Emergency Department.

Age – testing younger patients unnecessarily increases costs.

Comfort Measures Only – It is inappropriate to treat such patients beyond those measures necessary for comfort.

Clinical Trial – Rendering treatments not within the parameters of the trial will render invalid trial results of increase the length of the trial as “replacement” patients are recruited.

Bone Mineral Density Testing Performed in the Prior 12 Months, On FDA-Approved Pharmacotherapy for Osteoporosis Prior to Fracture Those patients already diagnosed with low bone mass have been evaluated previously; duplicative testing or treatment is not warranted.

Since the cost to abstract all data for measure 01 ranges from $2.81 to $6.86, the cost of abstracting for exclusions is a small fraction of those numbers and is minimal. The value of abstracting these data elements outweighs the burden and cost of abstraction.

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

**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?**

X☐ **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**.   
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*)

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

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

*Provide the statistical results from testing the approach to controlling for differences in patient characteristics (case mix) below*.  
***if stratified, skip to*** [***2b4.9***](#question2b49)

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

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

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

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

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

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

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

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

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

Descriptive statistics for the performance measure scores for all tested entities were constructed. These statistics were the mean, standard deviation, median, minimum, and maximum scores.

**2b5.2. What were the statistical results from testing the ability to identify statistically significant and/or clinically/practically meaningful differences in performance measure scores across measured entities?** (e.g., *number and percentage of entities with scores that were statistically significantly different from mean or some benchmark, different from expected; how was meaningful difference defined*)  
Meaningful difference was defined as a significant spread (>20%) between minimum and maximum scores or a significant spread between median and minimum or median and maximum scores.

Results for OAF-02 were: Mean: 7.4%

Standard Deviation: 18.9%

Median: 3.8%

Minimum: 0.0%

Maximum: 63.6%

**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?*)  
We interpret these results as showing a significant spread between both median and maximum scores and between minimum and maximum scores. These are indicative of statistically significant differences in performance. In addition, with a very low median of 3.8%, there is significant opportunity for improvement.

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

**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 datasources/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*)

**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*)