**National Quality Forum—Measure Testing (subcriteria 2a2, 2b1-2b6)**

**Measure Number** (*if previously endorsed*)**:** 0046

**Measure Title**: Screening for Osteoporosis for Women 65-85 Years of Age

**Date of Submission**: 4/9/2018

**Type of Measure:**

|  |  |
| --- | --- |
| Outcome (*including PRO-PM*) | Composite – ***STOP – use composite testing form*** |
| Intermediate Clinical Outcome | Cost/resource |
| Process *(including Appropriate Use)* | 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, 2b1, 2b2, and 2b4 must be completed.** * **For outcome and resource use measures**, section **2b3** also must be completed. * If specified for **multiple data sources/sets of specificaitons** (e.g., claims and EHRs), section **2b5** 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 (2b1-2b6) 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 25 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). * For information on the most updated guidance on how to address social risk factors variables and testing in this form refer to the release notes for version 7.1 of the Measure Testing Attachment. |

|  |
| --- |
| **Note:** The information provided in this form is intended to aid the Standing 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 **instrument-based measures** (including PRO-PMs) **and composite performance measures**, reliability should be demonstrated for the computed performance score.  **2b1.** **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 **instrument-based measures (including PRO-PMs) and composite performance measures**, validity should be demonstrated for the computed performance score.    **2b2.** **Exclusions** are supported by the clinical evidence and are of sufficient frequency to warrant inclusion in the specifications of the measure; [**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)  **2b3.** **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 (including clinical and social risk factors) that influence the measured outcome 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.   **2b4.** 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.  **2b5.** **If multiple data sources/methods are specified, there is demonstration they produce comparable results**.  **2b6.** 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. The degree of consensus and any areas of disagreement must be provided/discussed.  **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.** 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.17*)** | **Measure Tested with Data From:** |
| abstracted from paper record | abstracted from paper record |
| claims | claims |
| registry | registry |
| abstracted from electronic health record | abstracted from electronic health record |
| eMeasure (HQMF) implemented in EHRs | eMeasure (HQMF) implemented in EHRs |
| other: Click here to describe | other: Click here to describe |

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

**1.3. What are the dates of the data used in testing**? Click here to enter date range

2014 Submission:

Sample 1: Testing of data element reliability was performed during field testing in 2009.

Sample 2: Testing of performance variability was performed using 2012 performance data from the Physician Quality Reporting System.

**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.20*)** | **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*)

2014 Submission:

Sample 1: This measure was tested for data element reliability using field test data. To identify clinics for field testing, the American Academy of Orthopedic Surgeons (AAOS) posted an announcement online and also identified practices that were known through their previous work with the AAOS. Of the thirteen clinics who expressed an interest in the field-testing, two were chosen to participate. These two sites were chosen based on having participated in the 2009 Physician Quality Reporting Initiative (PQRI) program with additional consideration given to balancing practice size, location, and use of an EHR or paper medical record. One site was located in New Mexico and one was located in South Carolina.

Sample 2: This measure is used in the Physician Quality Reporting System (PQRS) as a performance measure for eligible professionals. 2012 performance data from PQRS was used to examine the variation in performance for this measure. The number of providers submitting data for this measure in 2012 was 35,079.

**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*)   
2014 Submission:

Sample 1: Desired sample size for testing was calculated for this measure with 0.80 power, 0.05 significance, and testing for a kappa of substantial agreement (0.8) versus moderate agreement (0.4). Expected performance was conservatively assumed at 0.5. Based on these assumptions and calculations, the minimum number of patients needed for the sample was 28. Both sites included 30 patients in their samples.

Sample 2: The number of patients eligible to be reported on in PQRS for 2012 was 13,339,356.

**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**.

2014 Submission:

Sample 1 was used to test data element reliability. Sample 2 was used to demonstrate performance variation.

**1.8** **What were the social risk factors that were available and analyzed**? For example, patient-reported data (e.g., income, education, language), proxy variables when social risk data are not collected from each patient (e.g. census tract), or patient community characteristics (e.g. percent vacant housing, crime rate) which do not have to be a proxy for patient-level data.

**2018 Submission:**

We did not analyze performance by social risk factors.

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

**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*)

2014 Submission:

**Critical data element reliability:** Reliability was tested by assessing whether two abstractors, reviewing the same full medical (including both inpatient and outpatient notes), would come to the same conclusion as to the patient meeting the measure, not meeting the measure, or qualifying as an exception. Two abstractors independently assessed whether patients met numerator inclusion criteria for each case that met denominator inclusion criteria. Following the data abstraction, the mismatches were tallied. Agreement between abstractors was measured using the kappa statistic (a measure of agreement adjusted for agreement that can occur by chance).

**2a2.3. For each level of testing checked above, what were the statistical results from reliability testing**? (e*.g., percent agreement and kappa for the critical data elements; distribution of reliability statistics from a signal-to-noise analysis*)  
2014 Submission:

**Denominator:** Agreement between the two independent reviewers was 100% for the denominator data element. The Kappa was undefined for this data element as both abstractors agreed that all of the cases met denominator eligibility criteria.

**Exceptions:** Agreement between the two independent reviewers was 100% for the exception data element. The Kappa was undefined for this data element as both abstractors agreed that none of the cases should be excluded.

**Numerator:** Agreement between the two reviewers was 90% with agreement that the numerator criteria was met in 19/30 cases and not met in 8/30 cases. The reviewers disagreed about 3/30 cases where one reviewer found evidence that the numerator criteria was met and one review did not find evidence in the medical record that numerator criteria was met. The abstractors then reconciled the mismatches through an adjudication process and determined 22/30 cases met numerator criteria. A Kappa statistic was calculated to demonstrate the degree of agreement adjusted for chance (K=0.77; 95% CI: 0.63-1.00).

Table 1 below displays the overall agreement for the all the measure components combined. Concordance between the abstractors is 90% with moderate agreement above what would be expected (K=0.77; 95% CI 0.53-1.00).

Table 1: Inter-rater reliability of measure components

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | | Reviewer B | | | |  |
| Not Study Eligible | Not Met | Met | Exception | Total |
| Reviewer A | Not Study Eligible | 0 | 0 | 0 | 0 | 0 |
| Not Met | 0 | 8 | 0 | 0 | 8 |
| Met | 0 | 3 | 19 | 0 | 22 |
| Exception | 0 | 0 | 0 | 0 | 0 |
| Total | 0 | 11 | 19 | 0 | 30 |

“Not Study Eligible” means that the denominator criteria were not met.

“Not Met” means denominator criteria were met, numerator criteria were not met and exceptions (exclusions) did not apply.

“Met” means denominator criteria were met and numerator criteria were met.

“Exception” means denominator criteria were met, numerator criteria were not met and exceptions applied.

|  |  |
| --- | --- |
| Kappa Coefficient | 0.77 |
| Kappa LL (95% Confidence Interval) | 0.53 |
| Kappa UL (95% Confidence Interval) | 1.00 |
| Observed Agreement Rate | 0.90 |
| Expected Agreement Rate | 0.56 |

**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?*)  
2014 Submission:

Interpretation of data element reliability testing: The below scale was used in the field test to interpret the kappa score. The numerator had a kappa score of .77, which indicates that there was substantial agreement that the two abstractors came to the same conclusion as to patients who met the numerator. This suggests the measure elements can be reliably abstracted from medical records.

Kappa Strength of Agreement

0.00 Poor

0.01 – 0.20 Slight

0.21 – 0.40 Fair

0.41 – 0.60 Moderate

0.61 – 0.80 Substantial

0.81 – 0.99 Almost perfect

Landis, J.R. and Koch, G. G. (1977) "The measurement of observer agreement for categorical data" in Biometrics. Vol. 33, pp. 159—174

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

**2b1. VALIDITY TESTING**

**2b1.1. 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*) **NOTE**: Empirical validity testing is expected at time of maintenance review; if not possible, justification is required.

**2b1.2. 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)*  
**2018 Submission:**

There are no updates to the validity testing for this measure since the last submission. The only publicly available data for this measure are from reporting in the CMS Quality Payment Program, however these data are not constructed in a way that allows NCQA to test empirical validity of the measure.

2014 Submission:

**Critical Data Element Validity:** The testing conducted for this measure by the AMA/Physician Consortium for Performance Improvement (PCPI) is described above under “Reliability.” This testing demonstrates inter-rater reliability of two reviewers using the same measure specification to draw conclusions from the same “gold-standard” data source (e.g. medical record). Reliability testing demonstrated that two independent reviewers looking at the same full medical record had high agreement on every data element and the overall performance measure score. We believe this testing demonstrates not only data element reliability but also validity, that is to say the accuracy of the measure specification to identify all data elements from the medical record.

**Assessment of face validity:** This measure was also evaluated for face-validity by the AMA/PCPI which oversees the measure development process of clinically relevant physician-level performance measures. To assess the face validity of measures, PCPI follows a standardized process for measure development which includes:

* Convening cross-specialty, multidisciplinary work groups to assess the face and content validity of each measure. The groups establish the measure’s ability to capture what it is designed to capture using a consensus process that consists of input from multiple stakeholders, including practicing physicians and experts with technical measure expertise.
* Review of the evidence, gaps in care and potential for impact of the measure:
  + Consider existing guideline recommendations and the strength of evidence
  + Consider gaps in care, variation, cost and frequency data
* Posting the draft measure for a 30-day public comment period. The PCPI solicits feedback from PCPI members, quality improvement collaboratives, providers, consumers, public/private purchasers and others with an interest in the measure.
* The PCPI work group reviews comments received, revises and modifies the draft performance measures as deemed appropriate by the work group. The public comments and responses are posted to the PCPI website as part of the voting process.
* Final vote by PCPI members eligible to vote. The PCPI encourages all voting member organizations to vote so the required quorum is met.

**2b1.3. What were the statistical results from validity testing**? (*e.g., correlation; t-test*)  
2014 Submission:

**Results of face validity assessment:** This measure was reviewed and developed by a joint work group that included experts in osteoporosis treatment as well as representatives from the following organizations: American Academy of Family Physicians; American Academy of Orthopaedic Surgeons; American Association of Clinical Endocrinologists; American College of Rheumatology; The Endocrine Society; American Medical Association; National Osteoporosis Foundation; National Committee for Quality Assurance; and The Joint Commission. The joint work group members came to consensus on the final recommended specification for this measure in October 2006. See section Ad. 1. Workgroup/Expert Panel Involved in Measure Development for a list of participants of the Osteoporosis Work Group.

**2b1.4. 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?*)  
2014 Submission:

**Interpretation of face validity assessment:** These results indicate that the multiple experts and stakeholders concluded with good agreement that the measure as specified is measuring what it intends to measure and that the results of the measurement allow users to make the correct conclusions about the quality of care that is provided and will accurately differentiate quality across providers.

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

**2b2. EXCLUSIONS ANALYSIS**

**NA**  **no exclusions — *skip to section*** [***2b3***](#section2b4)

**2b2.1. 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*)

2018 Submission:

The exclusion for this measure is based on clearly specified codes that indicate the patient received hospice services during the measurement period. While this code has not been specifically tested in the context of this measure, it is considered valid for identifying patients who receive hospice services. This measure does not allow for exclusions for patient refusal, provider refusal, or un-specified reasons.

2014 Submission:

N/A – no exclusions

**2b2.2. 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*)  
2018 Submission:

NA

2014 Submission:

N/A – no exclusions

**2b2.3. 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*)  
2018 Submission:

NA

2014 Submission:

N/A – no exclusions

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

**2b3. 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*** [***2b4***](#section2b5)***.***

**2b3.1. 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

**2b3.1.1 If using a statistical risk model, provide detailed risk model specifications, including the risk model method, risk factors, coefficients, equations, codes with descriptors, and definitions.**

**2b3.2. If an outcome or resource use component 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**.

**2b3.3a. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors (clinical factors or social risk 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*) **Also discuss any “ordering” of risk factor inclusion**; for example, are social risk factors added after all clinical factors?

**2b3.3b. How was the conceptual model of how social risk impacts this outcome developed? Please check all that apply:**

**Published literature**

**Internal data analysis**

**Other (please describe)**

**2b3.4a. What were the statistical results of the analyses used to select risk factors?**

**2b3.4b. Describe the analyses and interpretation resulting in the decision to select social risk factors** *(e.g. prevalence of the factor across measured entities, empirical association with the outcome, contribution of unique variation in the outcome, assessment of between-unit effects and within-unit effects.)* **Also describe the impact of adjusting for social risk (or not) on providers at high or low extremes of risk.**

**2b3.5. 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*** [***2b3.9***](#question2b49)

**2b3.6. Statistical Risk Model Discrimination Statistics** (*e.g., c-statistic, R-squared*)**:**

**2b3.7. Statistical Risk Model Calibration Statistics** (*e.g., Hosmer-Lemeshow statistic*):

**2b3.8. Statistical Risk Model Calibration – Risk decile plots or calibration curves**:

**2b3.9. Results of Risk Stratification Analysis**:

**2b3.10. 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*)

**2b3.11.** **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*)

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

**2b4. IDENTIFICATION OF STATISTICALLY SIGNIFICANT & MEANINGFUL DIFFERENCES IN PERFORMANCE**

**2b4.1. 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)*   
2014 Submission:

To demonstrate meaningful differences in performance, NCQA calculates an inter-quartile range (IQR). The IQR provides a measure of the dispersion of performance. The IQR can be interpreted as the difference between the 25th and 75th percentile on a measure.

**2b4.2. 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*)  
2014 Submission:

2012 Variation in Performance across Providers

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Mean Rate | EP | 10th | 25th | 50th | 75th | 90th | IQR |
| 58.7% | 326,372 | 0.00% | 22.7% | 64.3% | 100.0% | 100.0% | 77.3 |

EP: Number of patients meeting denominator criteria across all providers submitting data to the Physician Quality Reporting System on this measure

IQR: Interquartile range

**2b4.3. 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?*)  
2014 Submission:

The results above indicate there is a large gap in performance between providers at the 25th and 75th percentiles. This demonstrates a large variation in performance and significant room for improvement on this measure for many providers. It should be noted that performance data from the PQRS program does not reflect performance system wide because physicians have the option to report. We look forward to more detailed performance reports from PQRS that may demonstrate longitudinal provider-specific performance improvements.

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

**2b5. COMPARABILITY OF PERFORMANCE SCORES WHEN MORE THAN ONE SET OF SPECIFICATIONS**

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

**Note***: This item is directed to measures that are risk-adjusted (with or without social risk factors)* ***OR*** *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).* ***Comparability is not required when comparing performance scores with and without social risk factors in the risk adjustment model. However, if comparability is not demonstrated for measures with more than one set of specifications/instructions, the different specifications (e.g., for medical records vs. claims) should be submitted as separate measures.***

**2b5.1. Describe the method of testing conducted to compare 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*)

**2b5.2. 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*)

**2b5.3. What is your interpretation of the results in terms of the differences in 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*)

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

**2b6. MISSING DATA ANALYSIS AND MINIMIZING BIAS**

**2b6.1. 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*)  
2014 Submission:

This measure is collected with a complete sample through medical record review, there is no missing data on this measure.

**2b6.2. 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*)  
2014 Submission:

This measure is collected with a complete sample through medical record review, there is no missing data on this measure.

**2b6.3. 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*)

2014 Submission:

This measure is collected with a complete sample through medical record review, there is no missing data on this measure.