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

**Measure Number** (*if previously endorsed*)**:** 0053

**Measure Title**: Osteoporosis Management in Women Who Had a Fracture

**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*).

2014 Submission:

N/A

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

2014 Submission:

Sample 1: January 1 to December 31, 2012.

Sample 2: July 2000 through December 2001

**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 reliability and meaningful difference in performance at the plan level using data from all Medicare health plans submitting HEDIS data for measurement year 2012. The plans were nationally representative and included 235 HMO plans and 112 PPO plans. The plans varied in size from a minimum of 30 eligible patients to over 6,441 within a single plan.

**Sample 2:** This measure was originally field tested in a sample of 5 health plans. The five plans were geographically diverse and included both HMOs and PPOs.

**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:** In 2012, HEDIS measures covered 8.7 million Medicare beneficiaries. Data is summarized at the health plan level for all Medicare plans submitting data for this measure for 2012. Patients included in the HEDIS data include a diverse representation of ages, race and diagnoses. The table below shows the average number of eligible patients per health plan and the standard deviation of that average across health plans.

**Table 1: Sample 1 Average Eligible Population per Health Plan.**

|  |  |  |  |
| --- | --- | --- | --- |
| Product Type | Number of Plans | Average number of eligible patients per plan | Standard Deviation |
| Medicare | 347 | 372 | 625 |

**Sample 2:** The sample from the field test conducted in five health plans included all women who experienced a fracture between July 2000 and June 2001. Table 2 below shows the number of women who experienced a fracture in each health plan by age.

**Table 2: Sample 2 Eligible Population in each Field Test Health Plan**

|  |  |  |
| --- | --- | --- |
|  | Age | |
| Health Plan | 67 and over | 50 to 66 |
| Plan A | 202 | 207 |
| Plan B | 796 | 390 |
| Plan C | 613 | 201 |
| Plan D | 703 | 175 |
| Plan E | 876 | 199 |
| **Total** | **3,190** | **1,172** |

**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 demonstrate reliability (beta-binomial calculation), construct validity (correlation analysis) and meaningful difference in performance.

Sample 2 was used to field test the measure, test item-level validity and exclusions.

Plan-level validity was also demonstrated through a systematic assessment of face validity. This measure was systematically evaluated for face validity with four panels of experts:

* The Osteoporosis Advisory Workgroup included 5 experts in geriatrics, endocrinology, and osteoporosis.
* The Geriatric MAP included 13 experts in geriatrics, including representation by consumers, health plans, health care providers and policy makers.
* The Technical Measurement Advisory Panel includes 14 members, including representation by health plans methodologists, clinicians and HEDIS auditors.
* NCQA’s Committee on Performance Measurement (CPM) oversees the evolution of the measurement set and includes representation by purchasers, consumers, health plans, health care providers and policy makers. This panel is made up of 21 members. The CPM is organized and managed by NCQA and reports to the NCQA Board of Directors and is responsible for advising NCQA staff on the development and maintenance of performance measures. CPM members reflect the diversity of constituencies that performance measurement serves; some bring other perspectives and additional expertise in quality management and the science of measurement.

Per NQF instructions we have described the composition of the expert panels which assessed face validity for this measure. See Additional Information: Ad.1. Workgroup/Expert Panel Involved in Measure Development for names and affiliation of expert panels.

**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:

Plan-level reliability testing of performance measure score: In order to assess measure precision in the context of the observed variability across accountable entities, we utilized the reliability estimate proposed by Adams (2009). The following is quoted from the tutorial which focused on provider-level assessment: “Reliability is a key metric of the suitability of a measure for [provider] profiling because it describes how well one can confidently distinguish the performance of one physician from another. Conceptually, it is the ratio of signal to noise. The signal in this case is the proportion of the variability in measured performance that can be explained by real differences in performance. There are three main drivers of reliability: sample size, differences between physicians, and measurement error. At the physician level, sample size can be increased by increasing the number of patients in the physician’s data as well as increasing the number of measures per patient.” This approach is also relevant to health plans and other accountable entities.

Adams’ approach uses a Beta-binomial model to estimate reliability; this model provides a better fit when estimating the reliability of simple pass/fail rate measures as is the case with most HEDIS® measures. The beta-binomial approach accounts for the non-normal distribution of performance within and across accountable entities. Reliability scores vary from 0.0 to 1.0. A score of zero implies that all variation is attributed to measurement error (noise or the individual accountable entity variance) whereas a reliability of 1.0 implies that all variation is caused by a real difference in performance (across accountable entities).

Adams, J. L. The Reliability of Provider Profiling: A Tutorial. Santa Monica, California: RAND Corporation. TR-653-NCQA, 2009

**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:

**Results of reliability testing of performance measure score:** The table 3 below shows the results of the reliability testing of the performance measurement score in 2012.

**Table 3: Reliability in Medicare Plans in 2012**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| # of plans | Overall Reliability Score | 10th percentile | 25th percentile | 50th percentile | 75th percentile | 90th percentile |
| 347 | .92 | .81 | .89 | .95 | .97 | .99 |

**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 measure score reliability testing:** Reliability scores can vary from 0.0 to 1.0. A score of zero implies that all variation is attributed to measurement error (noise) whereas a reliability of 1.0 implies that all variation is caused by a real difference in performance (signal). Generally, a minimum reliability score of 0.7 is used to indicate sufficient signal strength to discriminate performance between accountable entities. The testing suggests this measure has very good reliability. The 10-90th percentile distribution of health plan level-reliability for this measure show nearly all health plans met or exceeded the minimally accepted threshold of 0.7, and the majority of plans exceeded 0.9. Strong reliability is demonstrated with the majority of variance attributed to signal and not to noise.

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

**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)*  
2014 Submission:

**Method of testing critical data element validity:** To test the validity of plan administrative data for computing this measure, participating field test plans (Sample 2) selected a random sample of 100 patients from their administrative data file and reviewed their primary care physician medical records. This data was used as a gold standard to verify the completeness and accuracy of the administrative data concerning the date and type of fracture, clinical exclusions, and dates and types of treatment. Given the small sample size and high level of agreement (see 2b.2.3), no statistical test of agreement was performed.

**Method of testing empirical validity:** We tested for construct validity by exploring whether performance for this measure was correlated with a similar measure, Osteoporosis Testing in Older Women, in the most recent year of available HEDIS data (Sample 1). This measure assesses the proportion of women who report having ever received a bone mineral density test to check for osteoporosis. The measures focus on the same disorder, osteoporosis, in different populations. We hypothesized that these two measures would be positively correlated (i.e. plans that have high rates of performance for management of osteoporosis will also have high rates of performance for screening of osteoporosis.) To test this correlation we used a Pearson correlation test. This test estimates the strength of the linear association between two continuous variables; the magnitude of correlation ranges from -1 and +1. A value of 1 indicates a perfect linear dependence in which increasing values on one variable is associated with increasing values of the second variable. A value of 0 indicates no linear association. A value of -1 indicates a perfect linear relationship in which increasing values of the first variable is associated with decreasing values of the second variable.

**2018 Submission:**

**Method of assessing face validity:** We describe below NCQA’s process for both measure development, and maintenance, which includes substantial feedback from 10 standing expert panels and 16 standing Measurement Advisory Panels, review and voting by our Committee on Performance Measurement and NCQA’s Board of Directors. In addition, all new measures and measures undergoing significant revision are included in our annual HEDIS 30-day public comment period, which on average receives over 800 distinct comments from the field including organizations that are measured by NCQA, providers, patients, policy makers and advocates. NCQA refines our measures continuously through feedback received from our Policy Clarification (PCS) Web Portal, which on average receives and responds to over 3,000 inquiries each year. All HEDIS measures are audited by certified firms according to standards, policies and procedures outlined in HEDIS Volume 7. Combined, these processes which NCQA has used for over 25 years assures that measures we use are valid.

NCQA has identified and refined measure management into a standardized process called the HEDIS measure life cycle for all plan-level HEDIS measures.

STEP 1: NCQA staff identifies areas of interest or gaps in care. Measurement Advisory Panels (MAPs) participate in this process. Once topics are identified, a literature review is conducted to find supporting documentation on their importance, scientific soundness and feasibility. This information is gathered into a work-up format. The work-up is vetted by NCQA’s MAPs, the Technical Measurement Advisory Panel (TMAP) and the Committee on Performance Measurement (CPM) as well as other panels as necessary.

STEP 2: Development ensures that measures are fully defined and tested before the organization collects them. MAPs participate in this process by helping identify the best measures for assessing health care performance in clinical areas identified in the topic selection phase. Development includes the following tasks: (1) Prepare a detailed conceptual and operational work-up that includes a testing proposal and (2) Collaborate with health plans to conduct field-tests that assess the feasibility and validity of potential measures. The CPM uses testing results and proposed final specifications to determine if the measure will move forward to Public Comment.

STEP 3: Public Comment is a 30-day period of review that allows interested parties to offer feedback to NCQA and the CPM about new measures or about changes to existing measures.   
NCQA MAPs and technical panels consider all comments and advise NCQA staff on appropriate recommendations brought to the CPM. The CPM reviews all comments before making a final decision about Public Comment measures. New measures and changes to existing measures approved by the CPM will be included in the next HEDIS year and reported as first-year measures.

STEP 4: First-year data collection requires organizations to collect, be audited on and report these measures, but results are not publicly reported in the first year and are not included in NCQA’s State of Health Care Quality, Quality Compass or in accreditation scoring. The first-year distinction guarantees that a measure can be effectively collected, reported and audited before it is used for public accountability or accreditation. This is not testing—the measure was already tested as part of its development—rather, it ensures that there are no unforeseen problems when the measure is implemented in the real world. NCQA’s experience is that the first year of large-scale data collection often reveals unanticipated issues. After collection, reporting and auditing on a one-year introductory basis, NCQA conducts a detailed evaluation of first-year data. The CPM uses evaluation results to decide whether the measure should become publicly reportable or whether it needs further modifications.

STEP 5: Public reporting is based on the first-year measure evaluation results. If the measure is approved, it will be publicly reported and may be used for scoring in accreditation.

STEP 6: Evaluation is the ongoing review of a measure’s performance and recommendations for its modification or retirement. Every measure is reviewed for reevaluation at least every three years. NCQA staff continually monitors the performance of publicly reported measures. Statistical analysis, audit result review and user comments through NCQA’s Policy Clarification Support portal contribute to measure refinement during re-evaluation. Information derived from analyzing the performance of existing measures is used to improve development of the next generation of measures.

Each year, NCQA prioritizes measures for re-evaluation and selected measures are researched for changes in clinical guidelines or in the health care delivery systems, and the results from previous years are analyzed. Measure work-ups are updated with new information gathered from the literature review, and the appropriate MAPs review the work-ups and the previous year’s data. If necessary, the measure specification may be updated or the measure may be recommended for retirement. The CPM reviews recommendations from the evaluation process and approves or rejects the recommendation. If approved, the change is included in the next year’s HEDIS Volume 2.

**ICD-10 conversion:** Goal was to convert this measure to a new code set, fully consistent with the intent of the original measure.

Steps in ICD-9 to ICD-10 Conversion Process

1. NCQA staff identify ICD-10 codes to be considered based on ICD-9 codes currently in measure. Use GEM to identify ICD-10 codes that map to ICD-9 codes. Review GEM mapping in both directions (ICD-9 to ICD-10 and ICD-10 to ICD-9) to identify potential trending issues.
2. NCQA staff identify additional codes (not identified by GEM mapping step) that should be considered. Using ICD-10 tabular list and ICD-10 Index, search by diagnosis or procedure name for appropriate codes.
3. NCQA HEDIS Expert Coding Panel review NCQA staff recommendations and provide feedback.
4. As needed, NCQA Measurement Advisory Panels perform clinical review. Due to increased specificity in ICD-10, new codes and definitions require review to confirm the diagnosis or procedure is intended to be included in the scope of the measure. Not all ICD-10 recommendations are reviewed by NCQA MAP; MAP review items are identified during staff conversion or by HEDIS Expert Coding Panel.
5. Post ICD-10 code recommendations for public review and comment.
6. Reconcile public comments. Obtain additional feedback from HEDIS Expert Coding Panel and MAPs as needed.
7. NCQA staff finalize ICD-10 code recommendations.

*Tools Used to Identify/Map to ICD-10:*

All tools used for mapping/code identification from CMS ICD-10 website (<http://www.cms.gov/Medicare/Coding/ICD10/2012-ICD-10-CM-and-GEMs.html>).

GEM, ICD-10 Guidelines, ICD-10-CM Tabular List of Diseases and Injuries, ICD-10-PCS Tabular List.

*Expert Participation:*

The NCQA HEDIS Expert Coding Panel and NCQA’s Diabetes Expert Panel reviewed and provided feedback on staff recommendations. Names and credentials of the experts who served on these panels are listed under Additional Information, Ad. 1. Workgroup/Expert Panel Involved in Measure Development.

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

**Results critical data element validity test:** The results in table 4 and 5 below show the number of numerator and denominator events identified in each field test plan’s sample of 100 patients. The results demonstrate high agreement between medical records and administrative data.

**Table 4: Numerator and Denominator events as identified by medical record and administrative data for adults age 65 and Older**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Denominator | | Numerator | | Rate | |
| Health Plan | MR | Admin | MR | Admin | MR | Admin |
| A | 35 | 31 | 2 | 2 | 6% | 6% |
| B | 37 | 37 | 6 | 6 | 16% | 16% |
| C | 9 | 11 | 1 | 1 | 11% | 9% |
| D | 27 | 32 | 2 | 1 | 7% | 3% |
| E | 27 | 28 | 5 | 6 | 19% | 21% |
| Total | 135 | 139 | 16 | 16 | 12% | 12% |

**Table 5: Numerator and Denominator events as identified by medical record and administrative data for adults age 50-64**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Denominator | | Numerator | | Rate | |
| Health Plan | MR | Admin | MR | Admin | MR | Admin |
| A | 35 | 32 | 5 | 4 | 14% | 13% |
| B | 20 | 20 | 1 | 1 | 5% | 5% |
| C | 3 | 2 | 2 | 0 | 67% | -- |
| D | 1 | 1 | 0 | 0 | -- | -- |
| E | 8 | 6 | 2 | 2 | 25% | 33% |
| Total | 67 | 61 | 10 | 7 | 15% | 11% |

**Results of empirical validity test:**

The results in Table 6 indicated that for plan-level reporting this measure was significantly (p<.05) correlated with the Osteoporosis Testing measure (NQF #0037) in the direction that was hypothesized.

|  |  |  |  |
| --- | --- | --- | --- |
| **Table 6. Correlation between Osteoporosis Measures in Medicare Plans - 2012** | | | |
| Pearson Correlation Coefficient | | | |
|  | Osteoporosis Testing in Older Women | | |
| Osteoporosis Management in Women who have had a Fracture | R=0.27305 (R Statistic)  p<.0001 (significance) | | |
| Note: All correlations are significant at p<.05 | |  |  |  |

**Results of face validity assessment:**

* Step 1: This measure was developed in 2002 to address under-diagnosis and treatment of osteoporosis in women who had fragility fractures. NCQA, along with the Osteoporosis Technical Subgroup and the Geriatric Measurement Advisory Panel, worked together to assess the most appropriate management steps for women who had a fragility fracture.
* Step 2: The measure was written and field-tested in 2002. After reviewing field test results, the CPM recommended to send the measure to public comment with a majority vote in January 2003.
* Step 3: The measure was released for Public Comment in 2003 prior to publication in HEDIS. The CPM recommended moving this measure to first year data collection by a majority vote.
* Step 4: The measure was introduced in HEDIS 2004. Organizations reported the measures in the first year and the results were analyzed for public reporting in the following year. The CPM recommended moving this measure to public reporting with a majority vote.
* Step 5: The measure was re-evaluated in 2013 and reviewed by the Osteoporosis Workgroup and the Geriatric Measurement Advisory Panel. The measure was presented to the CPM in January 2014 and proposed changes to the measure were posted for public comment February-March 2014. The CPM approved the proposed changes to the measure in May 2014 with a majority vote. These changes will go forward for use in HEDIS 2015.
* Conclusion: The measure was deemed to have the desirable attributes of a HEDIS measure in 2003 (relevance, scientific soundness, and feasibility).

**Results of ICD-10 conversion:**

Summary of Stakeholder Comments Received

NCQA posted ICD-10 codes for public review and comment in March 2011 and March 2012. NCQA received comments from four organizations:

* Support recommendations.
* Questions about select codes.
* Recommended additional codes for consideration.

**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 data element validity testing:** The results demonstrate near perfect agreement between medical records and administrative data. On average, health plans identified slightly more denominator events using administrative data and slightly more numerator events using medical record data. However, these differences were minor. We interpret this to suggest the administrative data elements used in this measure are valid compared to a gold standard medical record source.

**Interpretation of empirical validity testing:**

Coefficients with absolute value of less than 0.3 are generally considered indicative of weak associations whereas absolute values of 0.3 or higher denote moderate to strong associations. The significance of a correlation coefficient is evaluated by testing the hypothesis that an observed coefficient calculated for the sample is different from zero. The resulting p-value indicates the probability of obtaining a difference at least as large as the one observed due to chance alone. We used a threshold of 0.05 to evaluate the test results. P-values less than this threshold imply that it is unlikely that a non-zero coefficient was observed due to chance alone. *The results confirmed the hypothesis that this measure is correlated with the Osteoporosis Testing in Older Women (NQF #0037), suggesting they represent the* same underlying construct of quality of care for osteoporosis. Although the association was weak, it was significantly more than zero. A strong correlation would not be expected in this case due to the different denominators of these two measures.

**2018 Submission:**

**Interpretation of face validity assessment:**

NCQA’s expert panels, our measurement advisory panels and our Committee on Performance Measurement agreed that *Osteoporosis Management in Women Who Had a Fracture* 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 health plans.

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

**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*)  
2014 Submission:

At the time of field test, only one exclusion in the measure was tested, exclusion for women with prior treatment for osteoporosis (treatment on or within the 12 months prior to the fracture). The aim of testing exclusions in the field test data was to determine how common exclusions are in the eligible patient population and the impact of these exclusions on denominator sizes and performance rates. Our results (detailed below) show differences in performance rates with and without exclusions and across data sources (administrative vs. medical record).

**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*)  
2014 Submission:

On average 34% of women 65+ who experienced a fracture met the exclusion criteria for current treatment (prescription for treatment in past 12 months); 51% of women 50-64 met the exclusion criteria. The application of the exclusion to the measure reduced rates by more than 60% for both age groups (see Table 7). At the time of the field test, Hormone Replacement Therapy (HRT) was considered a safe treatment for osteoporosis. More than half (64%) of women age 65+ who met the exclusion criteria were prescribed HRT. Almost all (92%) women 50-64 who met exclusion criteria were prescribed HRT.

**Table 7: Exclusion for Treatment for Osteoporosis in prior 12 months**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Age 65+ | | | Age 50-64 | | |
|  | Number with current or prior RX | Rate with exclusion | Rate without exclusion | Number with current or prior RX | Rate with exclusion | Rate without exclusion |
| Plan A | 115 | 11% | 36% | 234 | 11% | 49% |
| Plan B | 616 | 11% | 41% | 495 | 14% | 53% |
| Plan C | 328 | 12% | 31% | 162 | 13% | 43% |
| Plan D | 244 | 13% | 31% | 135 | 11% | 41% |
| Plan E | 383 | 14% | 33% | 178 | 18% | 49% |
| Total | 1686 | 13% | 35% | 1204 | 13% | 48% |

To determine the most appropriate data source for identifying exclusions, we compared medical records to administrative data in a sample of approximately 100 patients per plan. Across all three sites, the majority of exclusions could be identified through administrative data. Only 1% of records had an exclusion identified through medical record data alone (see Table 8).

**Table 8: Exclusion for Treatment for Osteoporosis by Data Source**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Plan Name | Patients | Admin Only | MR Only | Admin & MR | Neither |
| Plan A | 116 | 6% | 0% | 40% | 54% |
| Plan B | 110 | 0% | 0% | 48% | 52% |
| Plan C | 100 | 82% | 2% | 5% | 11% |
| Plan D | 100 | 59% | 0% | 3% | 38% |
| Plan E | 100 | 62% | 3% | 4% | 31% |
| Total | 526 | 40% | 1% | 21% | 38% |

Admin: Administrative Data

MR: Medical Record

**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*)  
2014 Submission:

It is important to exclude women who have a prescription or therapy for osteoporosis treatment in the past 12 months so that the measure is focused on women who are not already on treatment at the time of the fracture. The exclusion looks back 12 months to identify a prescription for osteoporosis treatment because some osteoporosis treatments can be effective for up to 12 months. The field test identified that excluding women who have a prior prescription to treat osteoporosis in the past 12 months significantly impacts the measure, reducing rates by more than 60%. The test also identified that administrative data was sufficient to identify this exclusion. Therefore, we determined this exclusion to be important and feasible.

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

**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. To determine if this difference is statistically significant, NCQA calculates an independent sample t-test of the performance difference between two randomly selected plans at the 25th and 75th percentile. The t-test method calculates a testing statistic based on the sample size, performance rate, and standardized error of each plan. The test statistic is then compared against a normal distribution. If the p value of the test statistic is less than .05, then the two plans’ performance is significantly different from each other. Using this method, we compared the performance rates of two randomly selected plans, one plan in the 25th percentile and another plan in the 75th percentile of performance. We used these two plans as examples of measured entities. However, the method can be used for comparison of any two measured entities.

**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:

**Table 9: Variation in Performance across Health Plans in HEDIS (2012 data)**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Avg. EP | Mean Rate | SD | 10th | 25th | 50th | 75th | 90th | IQR |
| Medicare Plans | 372 | 23.1 | 13.7 | 12.2 | 15.0 | 19.1 | 25.9 | 40.5 | 10.9 |

EP: Eligible Population, the average denominator size across plans submitting to HEDIS

IQR: Interquartile range

**Table 10: T-test between two randomly selected health plans in HEDIS (2012 data)**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Plan Rate (25th Percentile) | Plan Rate (75th Percentile) | P-Value |
| Medicare Plans | 12.1 | 36.8 | .00003 |

p-value: P-value of independent samples t-test comparing plans at the 25th percentile to plans at the 75th percentile

**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 10.9% gap in performance between the 25th and 75th performing plans (see Table 9). The difference between the 25th and 75th percentile is statistically significant (see Table 10). This gap represents on average 40 more patients receiving bone mineral density testing or osteoporosis treatment following a fracture in high performing Medicare plans compared to low performing plans (estimated from average health plan eligible population).

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

**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:

Plans collect this measure using all administrative data sources. NCQA’s audit process checks that plans’ measure calculations are not biased due to missing data.

**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:

Plans collect this measure using all administrative data sources. NCQA’s audit process checks that plans’ measure calculations are not biased due to missing data.

**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:

Plans collect this measure using all administrative data sources. NCQA’s audit process checks that plans’ measure calculations are not biased due to missing data.

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

**Measure Number** (*if previously endorsed*)**:** 0053

**Measure Title**: Osteoporosis Management in Women Who Had a Fracture

**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*).

2014 Submission:

N/A

**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 using 2009 medical record data.

**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 at the provider level 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 (PQI) 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:** Reporting at the provider-level for this measure is collected through the Physician Quality Reporting System (PQRS). In 2012, 1730 providers nationwide reported on this measure. This reflects 0.8% of eligible providers.

**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 sizes for testing were 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, unless performance information was available for the measure. Based on these assumptions and calculations, the minimum number of patients needed for the sample was 38. A total of 39 patient records were reviewed across the two sites.

**Sample 2:** The number of beneficiaries reported on thought this measure for 2012 was 11,284.

**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 critical data element reliability.

Sample 2 was used to test meaningful differences in measure performance.

**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; with reviewers agreeing 36/39 cases met the denominator criteria. The Kappa was 1.00, which indicates that there was perfect agreement that the two abstractors came to the same conclusion as to patients who met the denominator.

**Exceptions:** Agreement between the two independent reviewers was 96.2% for the exception data element. The reviewers disagreed about 1/36 cases where one reviewer found evidence of the exception criteria and one reviewer found no evidence of the exception criteria. The Kappa was .65 for this data element. The reviewers met to reconcile their differences and determined the exception criteria was met in 1 case.

**Numerator:** Agreement between the two reviewers was 83.3% with agreement that the numerator criteria was met in 4/36 cases and not met in 24/36 cases. The reviewers disagreed about 6/36 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. A Kappa statistic was calculated to demonstrate the degree of agreement adjusted for chance (K=0.47; 95% CI: 0.11-.83). The two reviewers met to reconcile their differences and determined 5/36 cases met the numerator criteria for a performance rate of 13.9%.

Table 1 below displays the overall inter-rater reliability for the all the measure components combined. Concordance between the abstractors is 82% with moderate agreement above what would be expected (κ 95% confidence interval 0.35-.88).

**Table 1: Inter-rater reliability of measure components**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | | Reviewer B | | | | |
| Not Study Eligible | Not Met | Met | Exclusion | Total |
| Reviewer A | Not Study Eligible | 3 | 0 | 0 | 0 | 3 |
| Not Met | 0 | 24 | 2 | 1 | 27 |
| Met | 0 | 4 | 4 | 0 | 8 |
| Exclusion | 0 | 0 | 0 | 1 | 1 |
| Total | 3 | 28 | 6 | 2 | 39 |

“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.

“Exclusion” means denominator criteria were met, numerator criteria were not met and exclusion applied.

|  |  |
| --- | --- |
| Kappa Coefficient | 0.61 |
| Kappa LL (95% Confidence Interval) | 0.35 |
| Kappa UL (95% Confidence Interval) | 0.88 |
| Observed Agreement Rate | 0.82 |
| Expected Agreement Rate | 0.54 |

**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 denominator had a kappa score of 1.00, which indicates that there was perfect agreement that the two abstractors came to the same conclusion as to patients who met the denominator. The numerator had a kappa score of .47, which indicates that there was moderate agreement that the two abstractors came to the same conclusion as to patients who met the numerator. The exceptions part of this measure had a kappa score of .65, which indicates that there was substantial agreement that the two abstractors came to the same conclusion as to patients who met the exception criteria. Across all data elements the kappa was 0.61 indicating substantial agreement between raters. 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:** The AMA-convened Physician Consortium for Performance Improvement (PCPI) 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 exclusions for this measure are based on clearly specified codes that indicate the patient received hospice services or resided long-term in an institutional setting during the measurement period, patients who received previous pharmacologic therapy to treat osteoporosis in the previous 12 months or patients who had a bone mineral density test in the two years prior to the fracture. While these codes have not been specifically tested in the context of this measure, they are considered valid for identifying patients who should be excluded from the measure. This measure does not allow for exclusions for patient refusal, provider refusal, or un-specified reasons.

**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

**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

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

**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:

Table 4: Variation in Performance across Providers (2012 data)

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Mean Rate | EP | 10th | 25th | 50th | 75th | 90th | IQR |
| 70.0 | 11,284 | 0.0 | 25.0 | 100.0 | 100.0 | 100.0 | 75.0 |

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.