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

**Measure Number** (*if previously endorsed*)**:** 0037

**Measure Title**: Osteoporosis Testing in Older Women

**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: Patient Reported Data/Survey | other: Patient Reported Data/Survey |

**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*).

N/A

**1.3. What are the dates of the data used in testing**?

**2018 Submission:**

Sample 3: 2016

**2014 submission:**

Sample 1: 2005

Sample 2: 2012

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

|  |  |
| --- | --- |
| **Measure Specified to Measure Performance of:**  **(*must be consistent with levels entered in item S.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*)

**2018 Submission:**

**Sample 3:** To test the incidence of the exclusion for individuals receiving hospice care (first incorporated into the measures in 2016), we used data from Medicare health plans submitting Health Outcome Survey data to be reported in HEDIS for measurement year 2016. The plans were nationally representative and included 463 PPO and HMO plans.

**2014 submission:**

**Sample 1:** To test data element reliability and validity, NCQA contracted with RTI to conduct four rounds of cognitive testing between January and May 2005 in Raleigh and Durham, North Carolina, and Waltham, Massachusetts. Six respondents in each round for a total of 24 completed interviews. There were two rounds of concept testing to identify which terms used to describe osteoporosis and osteoporosis testing were recognized and understood by respondents. Using the terms identified in Rounds 1 and 2, the osteoporosis survey question was then tested in Rounds 3 and 4.

**Sample 2:** This measure was tested for reliability, empirical validity, meaningful difference in performance and missing data using data from Medicare health plans submitting Health Outcome Survey data to be reported in HEDIS for measurement year 2012. The plans were nationally representative and included 495 PPO and HMO plans.

**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*)   
**2018 Submission:**

**Sample 3**: In 2016, this measure was collected from 302,404 survey responders (patients) from 463 health plans. A sample was drawn from each health plan’s population based on the plan size (500 to 1,200 patients per health plan are sampled depending on plan size).

**2014 submission:**

**Sample 1:** Four rounds of cognitive testing took place between January and May 2005. Six respondents were interviewed in each round for a total of 24 completed interviews. The first two rounds of testing were done to identify which terms used to describe osteoporosis and osteoporosis testing were recognized and understood by respondents. The survey question was then tested in rounds three and four and were based off of the terms that were identified in rounds one and two. Participants were recruited for each round of cognitive testing from senior centers, senior housing, physical therapy, wellness centers, physicians’ offices and by word of mouth. In addition, announcements were placed about the study in local newspapers. Respondents were also required to have seen a health care provider during the past year. For the fourth round of testing, women aged 65 and older who had been diagnosed with osteoporosis were recruited. Round four included two women 65-75 years of age and four women who were over the age of 75. Respondents in round four were also diverse on their level of education. One respondent had less than high school, one had some high school, three were high school graduates or had their GED, and one woman had a four-year college degree or more.

**Sample 2**: In 2012, this measure was collected from 297,974 survey responders (patients) from 495 health plans. A sample was drawn from each health plan’s population (1,200 beneficiaries per health plan sampled). Table 1 below lists the demographic characteristics of the 2012 cohort.

**Table 1: Demographic Characteristics of Sample 2 (Health Outcome Survey 2012 Cohort)**

|  |  | **%** |
| --- | --- | --- |
| Age | Under 65 | 15.5 |
| 65–69 | 25.4 |
| 70–74 | 22.1 |
| 75–79 | 16.3 |
| 80 and older | 20.8 |
| Gender | Male | 42.5 |
| Female | 57.5 |
| Race | Hispanic | 3.2 |
| North American Native | 0.3 |
| Asian | 2.2 |
| Black | 12.2 |
| White | 79.6 |
| Other | 2.1 |
| Unknown | 0.4 |

**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**.

**2018 Submission:**

Sample 3 was used to test the incidence of the exclusion for individuals receiving hospice care.

**2014 submission:**

Sample 1 was used to test item-level reliability and validity.

Sample 2 was used to test reliability, empirical validity, meaningful difference in performance, and missing data.

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:**

**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:**

**Table 2: Reliability in Medicare Plans in 2012**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| # of plans | Overall Reliability Score | 10th percentile | 25th percentile | 50th percentile | 75th percentile | 90th percentile |
| 495 | 0.995 | 0.994 | 0.995 | 0.996 | 0.997 | 0.997 |

**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 reliability output from HEDIS 2012 data shows high overall reliability with a mean individual reliability above .9 for all plans. The lowest individual reliability found among plans was .92, the highest individual reliability was .99. Reliability assesses the degree to which a measure produces stable and consistent results, therefore, there is a high degree of consistency in the results and the variability between plans is most likely due to the performance of plans. Eight plans did not have a denominator of >30 and were not included in the reliability analysis.

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

**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:** Cognitive testing is a process used routinely in determining the content validity of survey questions. Through cognitive testing, trained interviewers assess whether respondents understand survey questions, can recall the information being asked of them, and answer the questions correctly given their experiences. Cognitive interviewing involves asking volunteer respondents to answer the survey questions (either on paper or verbally) and then interviewing respondents about their answers to the questions. Interview protocols include specific questions about the respondent’s thought-process when answering the questions. Questions are tested in rounds to allow for revision to the survey questions and interview protocol between testing rounds. Testing was completed by a professional research team at RTI. The text below describes the specific testing protocol in greater detail:

There were four rounds of testing for this measure. The first two rounds of testing focused on concept testing with the goal of determining whether women were familiar with the term “osteoporosis,” the best term to use as a descriptor of osteoporosis, and the best term to use for osteoporosis testing. We also tested the effectiveness of adding a description of the osteoporosis test.

The following terms were tested in Rounds 1 and 2 (terms that did not test well were dropped for the second round of concept testing):

Osteoporosis (descriptors)

* Bone loss
* Weakening of the bones
* Thin bones
* Brittle bones

Osteoporosis testing

* Bone density test
* Bone scan
* Dexa-scan
* Densitometry
* Bone mineral test
* Bone ultrasound
* BMD test

**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.

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. Refer to What Makes a Measure “Desirable”? 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.

**Method of Testing Empirical Validity**: We tested for construct validity by exploring whether performance for this measure was correlated with a similar measure, Osteoporosis Management in Women Who Had a Fracture. This measure assesses the percentage of women who experienced a fracture and received either bone mineral density test or a prescription for an osteoporosis treatment. The intent of the Osteoporosis management measure is to assess a health plan’s performance at secondary prevention of osteoporosis related fracture. We specifically 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.

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

**Results of Critical Data Element Validity Testing:**Most women had heard the term “osteoporosis” before testing and were usually able to accurately describe it. Several descriptors were tested including “brittle bones,” “bone loss,” and weakening of the bones.” Although not unanimous, most women picked “brittle bones” as their top choice of a descriptor, but even those who did not pick this term as a top choice were still familiar with it and thought it was an accurate way to describe osteoporosis. We tested a long list of terms used to describe the osteoporosis test. Most of the terms were technical and not known by respondents (e.g., DXA-scan, densitometry). The term “bone density test” was the term most familiar to respondents. The effectiveness of adding a description of the test was also examined (“This test may have been done to your back or hip”). All of the respondents found this addition to be helpful.

**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 screening and treatment 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 (relevance, scientific soundness, and feasibility).

**Results of Construct Validity Testing**: The results in Table 1a indicated that the Osteoporosis Testing measure was significantly (p<.05) correlated with the Osteoporosis Management measure (NQF #0053) in the direction that was hypothesized.

|  |  |
| --- | --- |
| **Table 3: 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) |

**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 Critical Data Element Validity Testing**: Cognitive testing showed that the terms used in the measure are understandable and familiar to most women.

**Interpretation of Construct 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 greater 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*)  
**2018 Submission:**

The exclusion for this measure for individuals in hospice care is based on using a Hospice Status flag in the file that contains all the CMS Health Outcome Survey data submission. This measure does not allow for exclusions for patient refusal, provider refusal, or un-specified reasons. While we did not fully test this exclusion and its impact on measure performance, using data from measurement year 2016 we examined the total number and percent of individuals who were excluded from measure reporting based on the Hospice Status Flag.

**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:**

The total Health Outcome Survey quality reporting sample was 668,143 individuals. Of these, 4,677 (0.7%) individuals had the Hospice Status flag and were excluded from the reporting of this measure.

**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:**

The 0.7% incidence of those meeting the hospice exclusion criteria is in line with what we would expect to see given publicly available data published by CMS on the use of hospice services among Medicare Advantage beneficiaries. While we were not able to test the impact of the exclusion on performance measure score, excluding individuals in hospice care from getting bone mineral density tests to screen for osteoporosis makes clinical sense. The exclusion was implemented using data that was already collected and reported (the Hospice Status Flag) and therefore added no additional burden to measure reporting.

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

**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**.   
**2014 submission:**

This measure is a process measure collected from patient self-report. Although this measure is collected from patient self-report it is not a PRO-PM, as it does not assess a patient reported outcome. Therefore we do not risk-adjust the rates.

**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?  
**2014 submission:**

N/A

**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?  
2014 submission:**

N/A

**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.**

**2014 submission:**

N/A

**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*)  
**2014 submission:**

N/A

**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*)

**2014 submission:**

N/A

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

**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) for each indicator. 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 4: Variation in Performance across Medicare Health Plans in HEDIS (2012 data)**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Avg. | SD | 10th | 25th | 50th | 75th | 90th | IQR |
| Medicare Plans | 73.1 | 9.6 | 59.3 | 67.1 | 74.6 | 81.0 | 84.1 | 13.9 |

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

IQR: Interquartile range

**Table 5: T-test between two randomly selected health plans in HEDIS (2012 data)**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Plan Rate (25th Percentile) | Plan Rate (75th Percentile) | Z-score | P-Value |
| Medicare | 65.8 | 82.2 | 4.3 | <.05 |

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 13.9 percent gap in performance between the 25th and 75th performing plans. The difference between the 25th and 75th percentile is statistically significant.

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

**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 through the Health Outcomes Survey. Our analysis of missing data was done at two levels: survey-level and item-level.

**Survey Level Missing Data Analysis:** NCQA conducted an analysis in 2013 to investigate whether there were significant differences between responders, late responders and nonresponders to the Health Outcome Survey. Where data are collected in waves, such as with the HOS, there is the opportunity to estimate the potential impact of nonresponse by studying specific subsets of the responder pool. The classic concept of the continuum of resistance postulates that individuals who fail to respond to multiple survey attempts are increasingly more resistant to completing a survey and that the most difficult-to-reach respondents may be similar to nonrespondents (Halbesleen 2013). In the context of the HOS, members can be classified as “late responders” (time to survey completion exceeded the Cohort and administration-specific 90th percentile), “other responders” and “nonresponders.” Because within the concepts of wave analysis, late responders may be more similar to nonresponders, the late responders population can be compared with all other responders to estimate how different true nonresponders may be from responders. We estimated differences in member characteristics across multiple years in a sample of responders and nonresponders (only 2010 data is displayed below, results were across multiple years). The characteristics compared across populations included: CMS-Hierarchical Condition Categories (HCC) risk score (an administrative proxy for health status), age, gender, race, disability status, hospice status, institutionalization status, and end stage renal disease (ESRD) status. These characteristics were available in CMS administrative systems and were not obtained through the Health Outcomes Survey. Additional characteristics for late responders and on-time responders were drawn from the HOS survey: household income, home ownership, marital status, and education level. Given the large sample size, differences between responders, late responders and nonresponders were evaluated using effect size calculations (Cohen’s D for continuous variables and Cramer’s V for nominal variables.) Cramer’s V for nominal variables is used to examine the association of two values. The result is between 0 (no association) and +1 (complete association). Cohen’s D calculates the difference between two means divided by the standard deviation for the data. This formula is typically used to estimate needed samples sizes although here it was used to compare differences in means between samples. A lower score indicates the need for a larger sample size where as a higher score indicates a direct correlation between two means.

Halbesleben, J.R.B., and M.V. Whitman. 2013. Evaluating Survey Quality in Health Services Research: A Decision Framework for Assessing Nonresponse Bias. Health Serv Res.Jun;48(3):913-30.

**Item Level Missing Data Analysis:** To further understand the potential impact of missing data we calculated the rate of item-level missing data for survey responders in 2012. We calculated the average rate of missing data on the osteoporosis question, the distribution of missing data across health plans and the frequency of 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:**

**Survey Level Missing Data Analysis:**

In 2010 the response rate to the Health Outcome Survey was 63.3%, among responders 9.1% were “late responders” who responded (time to survey completion exceeded the Cohort and administration-specific 90th percentile.) Table 6 shows differences between responders, late responders and nonresponders in a sample from 2010, and the effect size of those differences. Due to sample size, nearly all differences in mean or proportion were significant, although most differences were small when evaluated in terms of effect sizes with the exception of race.

**Table 6: Characteristics of Survey Responders, Late Responders and Nonresponders**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Mean or Percentage  of Total** | | | **Significance and  Effect Size1** | | |
| **Responder** | **Non** | **Late** | **Other vs. Non** | **Late vs. Other** | **Late vs. Non** |
| Age2,3 | 72.9 | 71.4 | 71.5 |  |  |  |
| HCC2,3 | 1.1 | 1.2 | 1.1 |  |  |  |
| Male | 42.9 | 44.8 | 43.4 |  |  |  |
| Non-White | 19.0 | 27.5 | 23.5 | \*\* | \* |  |
| Disability | 14.1 | 19.7 | 19.2 | \* |  |  |
| Dual Eligibility | 19.9 | 29.3 | 25.0 | \*\* |  |  |
| Institutionalized | 0.7 | 3.4 | .4 | \*\* |  | \* |
| Hospice | 0.1 | 0.2 | .1 |  |  |  |
| ESRD | 0.1 | 0.1 | .1 |  |  |  |
| Household Income (<20,000) | 46.0 | 38.7 |  | \* |  |  |
| Not a Homeowner | 38.5 | 34.0 |  |  |  |  |
| Education Level—Less Than High School | 37.5 | 27.6 |  | \* |  |  |
| Marital Status—Not Married | 45.9 | 47.2 |  |  |  |  |

1 Effect size estimates for pairwise comparison of nominal variables for responders and nonresponders were based on Cramer's V. The following classification was used: 0 to <0.05 = no effect; 0.05 to <0.1 = weak effect (\*); 0.10 to <0.15 = moderate effect (\*\*); 0.15 to <0.25 = strong effect (\*\*\*); >0.25 = very strong effect (\*\*\*\*).

2 Effect size estimates for pairwise comparisons of continuous variable for responders and nonresponders were based on Cohen's D, whereas the overall effect of response group on means of the response variable were based on omega squared. For Cohen's D, the following classification was used: 0 to <0.2 = no effect; 0.2 to <0.5 = small effect (\*); 0.5 to <0.8 = medium effect (\*\*); >0.8 = large effect (\*\*\*).

3 Group means are displayed for continuous variables.

**Item Level Missing Data:**

Almost all health plans (96%) had less than 5% missing response to the osteoporosis item among survey responders (See Table 7). The average missing item rate across health plans was 2% (See Table 8).

**Table 7: Frequency of Missing Data for Osteoporosis Item (HEDIS 2012 data)**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Missing %** | **Frequency** | **Percent** | **Cumulative Freq** | **Cumulative %** |
| **< 5%** | 477 | 96.36 | 477 | 96.36 |
| **5% - <10%** | 12 | 2.42 | 489 | 98.79 |
| **10% - <15%** | 5 | 1.01 | 494 | 99.8 |
| **>= 15%** | 1 | 0.2 | 495 | 100 |

**Table 8: Distribution of Non-response of Osteoporosis Item Across Plans (HEDIS 2012 data)**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Description** | **N** | **Mean** | | **std** | **p10** | **p25** | **p50** | **p75** | **p90** |
| Missing | 495 | 2.17% | | 1.71% | 0.58% | 1.21% | 1.90% | 2.69% | 3.65% |
| N: number of plans | | |
| std: standard deviation | | |
| p: percentile | | |

**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:**

**Survey Level Missing Data Analysis:** In general, late responders tended to be similar to nonresponders on every variable except institutionalization.Responders and non-responders tended to be similar in terms of age, health (HCC score), and gender. While most differences between groups were small, there were moderate differences seen between responders and non-responders with regard to the percent of individuals who were non-white, had dual eligibility, or were institutionalized. Analysis of the effect size showed none of these differences to be large or strong. It is not surprising that individuals with dual eligibility, disability or in institutions are less likely to respond. This population likely has a higher rate of cognitive impairment. The dual eligible population is also more likely to be non-English speaking (the mailed survey is offered in English, Spanish, and Chinese). Overall, our measurement advisory panel did not feel these differences reflected significant non-response survey bias.

**Item Level Missing Data Analysis:** The overall frequency of missing data for the OTO questions was very low across plans. Over 96% of plans had 5% or fewer missing responses for the OTO question. Only one plan was missing data for the OTO question for more than 15% of their survey population. The distribution showed that on average across plans, survey responses were missing data on the OTO question 2.2% of the time. Based on this analysis, it is unlikely that missing data on this question would bias performance results.