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

**Measure Number** (*if previously endorsed*)**:** 2372

**Measure Title**: Breast Cancer Screening

**Date of Submission**: 4/16/2018

**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 decision-making) 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*).

N/A

**1.3. What are the dates of the data used in testing**? 2014 submission: 2010-2012; 2018 submission: 2016-2017

**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

MEASURE SCORE RELIABILITY TESTING

Measure score reliability was calculated from HEDIS data that included all plans submitting data to NCQA for HEDIS in 2017: 420 commercial plans, 431 Medicare plans, and 257 Medicaid plans. The plans were geographically diverse and varied in size.

CONSTRUCT VALIDITY TESTING

Measure score reliability was calculated from HEDIS data that included all plans submitting data to NCQA for HEDIS in 2016: 426 commercial plans and 405 Medicare plans. The plans were geographically diverse and varied in size.

SYSTEMATIC EVALUATION OF FACE VALIDITY: same as below

2014 Submission

MEASURE SCORE RELIABILITY TESTING

Measure score reliability was calculated from HEDIS data that included all plans submitting data to NCQA for HEDIS: 419 commercial plans, 478 Medicare plans, and 168 Medicaid plans. The plans were geographically diverse and varied in size.

CONSTRUCT VALIDITY TESTING

Measure score reliability was calculated from HEDIS data that included all plans submitting data to NCQA for HEDIS: 419 commercial plans, 478 Medicare plans, and 168 Medicaid plans. The plans were geographically diverse and varied in size.

SYSTEMATIC EVALUATION OF FACE VALIDITY

This measure was tested for face validity with three panels of experts:

The Breast Cancer Screening Measurement Advisory Panel includes 9 experts in breast cancer care, including representation by consumers, health plans, health care providers, academia and policymakers.

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 HEDIS measurement set and includes representation by purchasers, consumers, health plans, health care providers and policy makers. The CPM is composed of 21 members, is organized and managed by NCQA, and reports to the NCQA Board of Directors. The CPM advises 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.

**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

Patient sample for measure score reliability testing

2017 data are stratified by product line (i.e. commercial, Medicaid, Medicare). Below is a description of the sample. It includes the number of health plans included in HEDIS data collection and the median eligible population for the measure across health plans.

|  |  |  |
| --- | --- | --- |
| Product Type | Number of Plans | Median number of eligible patients per plan |
| Commercial | 420 | 7,740 |
| Medicaid | 258 | 2,439 |
| Medicare | 431 | 1,890 |

Patient sample for construct validity testing

2016 data are stratified by product line (i.e. commercial, Medicaid, Medicare). Below is a description of the sample. It includes the number of health plans included in HEDIS data collection and the median eligible population for the measure across health plans.

|  |  |  |
| --- | --- | --- |
| Product Type | Number of Plans | Median number of eligible patients per plan |
| Commercial | 426 | 7,510 |
| Medicaid | 202 | 1,419 |
| Medicare | 405 | 2,297 |

2014 Submission

Patient sample for measure score reliability and validity testing

2013 Data are stratified by product line (i.e. commercial, Medicaid, Medicare). Below is a description of the sample. It includes the number of health plans included in HEDIS data collection and the median eligible population for the measure across health plans.

|  |  |  |
| --- | --- | --- |
| Product Type | Number of Plans | Median number of eligible patients per plan |
| Commercial HMO | 219 | 26,080 |
| Commercial PPO | 200 | 49,405 |
| Medicaid HMO | 165 | 4,065 |
| Medicare HMO | 330 | 2,948 |
| Medicare PPO | 148 | 2,202 |

**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

Same as below

2014 Submission

The same data were used for reliability and construct validity as described above.

In addition, validity was demonstrated through a systematic assessment of face validity. Per NQF instructions, we have described the composition of the expert panels that assessed face validity in the data sample questions above.

**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

For Medicare health plans, this measure was analyzed by low-income subsidy, dual eligibility and disability status, which served as proxies for lower socioeconomic status. These are available data elements for Medicare plans.

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

**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*)

2018 Submission

We assessed reliability of the measure in 2017 using the same methods specified below in the 2014 submission.

2014 Submission

METHODS FOR BETA-BINOMIAL RELIABILITY TESTING

The beta-binomial method (Adams, 2009) measures the proportion of total variation attributable to a health plan, which represents the “signal”. The beta-binomial model also estimates the proportion of variation attributable to measurement error for each plan, which represents “noise”. The reliability of the measure is represented as the ratio of signal to noise.

* A score of 0 indicates none of the variation (signal) is attributable to the plan
* A score of 1.0 indicates all of the variation (signal) is attributable to the plan
* A score of 0.7 or higher indicates adequate reliability to distinguish performance between two plans

PLAN-LEVEL RELIABILITY

The underlying formulas for the beta-binomial reliability can be adapted to construct a plan-specific estimate of reliability by substituting variation in the individual plan’s variation for the average plan’s variation. Thus, the reliability for some plans may be more or less than the overall reliability across plans.

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*)

2018 Submission

Beta-Binomial Statistic:

|  |  |  |
| --- | --- | --- |
| Commercial | Medicare | Medicaid |
| 0.998 | 0.997 | 0.993 |

2014 Submission

The reliability for the Breast Cancer Screening measure was estimated at 1.0 for commercial, 0.99 for Medicaid, and 0.99 for Medicare based on 419 commercial plans, 478 Medicare plans, and 168 Medicaid plans.

PLAN-LEVEL RELIABILITY

This table summarizes the variability of individual plan reliability. The reliability among the 10th percentile-plans was above 0.7, indicating high reliability for the majority of plans.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Overall Reliability | Median | 10th percentile, 90th percentile |
| Commercial | 0.99 | 1.00 | 0.97, 1.00 |
| Medicaid | 0.96 | 0.99 | 0.89, 1.00 |
| Medicare | 0.95 | 0.98 | 0.84, 1.00 |

**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?*)

2018 Submission

Interpretation of measure score reliability testing: Testing indicates the measure has very high reliability.

2014 Submission

Results indicate the measure has a strong signal to noise ratio, thus having sufficient signal strength to discriminate performance between accountable entities. Our results suggest the measure is highly reliable.

At the plan level, the vast majority of plans met or exceeded the minimally accepted threshold of 0.7, and the majority of plans exceeded 0.9.

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

**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

We assessed face validity of the measure in 2017, using the same methods specified below. Specifically, we assessed the implication of adding digital breast tomosynthesis as an acceptable breast cancer screening method to account for the use of this method by women with clinical indications. We also assessed construct validity of the measure using 2016 data and using the same methods specified below. Specifically, we assessed correlations between the Breast Cancer Screening and Colorectal Cancer Screening measures (commercial and Medicare plans).

2014 Submission

Method of Assessing Face Validity: 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. Clinical expert panels (MAPs—whose members are authorities on clinical priorities for measurement) 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 Measurement Advisory Panels (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 that allows interested parties to offer feedback to NCQA about new measures or changes to existing measures. NCQA MAPs and technical panels consider all comments and advise NCQA staff on recommendations brought to the CPM. The CPM reviews all comments before making a final decision about measures. New measures and changes to existing measures approved by the CPM are included in the next HEDIS year.

STEP 4: First-year data collection requires organizations to collect, be audited on and report these measures, but results are not publicly reported nor included in NCQA’s State of Health Care Quality, Quality Compass or in accreditation scoring. 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 during real-world implementation. 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 results from previous years are analyzed. Measure work-ups are updated, and the appropriate MAPs review the work-ups and data. If necessary, the measure specifications 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 construct validity

We tested for construct validity by exploring whether the measure was correlated with measures of quality hypothesized to be related. The Pearson correlation test is used to examine the association between the measures; the test estimates the strength of the linear association between two continuous variables and the magnitude of correlation ranges from -1 and +1, inclusive. A value of 1 indicates a perfect linear dependence in which increasing values on one variable are 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 are associated with decreasing values of the second variable. Coefficients with absolute values 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 it is unlikely that a non-zero coefficient was observed due to chance alone.

For the Breast Cancer Screening measure, we assessed correlations with Colorectal Cancer Screening (commercial and Medicare plans) and Cervical Cancer Screening (commercial and Medicaid plans). Our hypothesis was that these three measures would be positively correlated, as they assess secondary prevention services specific to cancer. We would expect plans that perform highly on Breast Cancer Screening to also perform highly on Colorectal Cancer Screening and Cervical Cancer Screening.

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

2018 Submission

Results of face validity assessment: Input from our multi-stakeholder measurement advisory panels and those submitting to public comment indicate the measure has face validity and supported adding digital breast tomosynthesis as a screening method.

Statistical results of construct validity testing: The results in Table 1a and Table 1b indicate that there is a strong, positive relationship between the Breast Cancer Screening measure and the Colorectal Cancer Screening measure. This relationship is statistically significant (p<0.0001).

**Table 1a. Correlations in Commercial Measures – 2016**

|  |  |
| --- | --- |
|  | **Pearson Correlation Coefficient** |
| Colorectal Cancer Screening |
| Breast Cancer Screening | 0.71 |

Note: p<0.0001

**Table 1b. Correlations in Medicare Measures – 2016**

|  |  |
| --- | --- |
|  | **Pearson Correlation Coefficient** |
| Colorectal Cancer Screening |
| Breast Cancer Screening | 0.72 |

Note: p<0.0001

2014 Submission

Face Validity: This measure was re-evaluated in 2012-2013. NCQA and the Breast Cancer Screening MAP worked together to assess the most appropriate ages and frequency for mammography screening using the 2009 US Preventive Services Task Force and other national guidelines. After reviewing the updated evidence and the recommendations from the MAP, the CPM recommended to send the measure to public comment with a majority vote. We received and responded to 340 comments on this measure, adjusting the measure as determined to be necessary, working with our advisory panels. The CPM recommended moving this measure into HEDIS with a majority vote.

Construct Validity: Pearson Correlation Coefficient results are shown in Tables 1-3.

Table 1. Correlation between Breast Cancer Screening, Colorectal Cancer Screening, and Cervical Cancer Screening, Commercial 2013

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Pearson Correlation Coefficients** | | |
| Breast Cancer Screening | Colorectal Cancer Screening | Cervical Cancer Screening |
| Breast Cancer Screening | 1 | 0.73 | 0.70 |
| Colorectal Cancer Screening |  | 1 | 0.59 |
| Cervical Cancer Screening |  |  | 1 |
| Note: All correlations are significant at p<0.05  Table 2. Correlation between Breast Cancer Screening and Cervical Cancer Screening, Medicaid 2013   |  |  |  | | --- | --- | --- | |  | **Pearson Correlation Coefficients** | | | Breast Cancer Screening | Cervical Cancer Screening | | Breast Cancer Screening | 1 | 0.56 | | Cervical Cancer Screening |  | 1 | | | | |

Note: All correlations are significant at p<0.05

Table 3. Correlation between Breast Cancer Screening and Colorectal Cancer Screening, Medicare 2013

|  |  |  |
| --- | --- | --- |
|  | **Pearson Correlation Coefficients** | |
| Breast Cancer Screening | Colorectal Cancer Screening |
| Breast Cancer Screening | 1 | 0.81 |
| Colorectal Cancer Screening |  | 1 |

Note: All correlations are significant at p<0.05

**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?*)

2018 Submission

Interpretation of systematic assessment of face validity: The measurement advisory panels showed good agreement that the measure as specified will accurately differentiate quality across providers. Our interpretation of these results is that this measure has sufficient face validity.

Interpretation of construct validity testing: The two measures had high correlation, which indicates the measure has good construct validity.

2014 Submission

FACE VALIDITY

Multiple NCQA panels concluded with good agreement that the measures as specified accurately to assess breast cancer screening in health plans. This measure meets the test for face validity.

CORRELATIONS

As hypothesized, Breast Cancer Screening was strongly positively correlated to the Colorectal Cancer Screening (0.73) and Cervical Cancer Screening (0.70) measures in commercial plans. Breast Cancer Screening was moderately positively correlated to the Cervical Cancer Screening (0.56) measure in Medicaid plans. Breast Cancer Screening was strongly positively correlated to the Colorectal Cancer Screening (0.81) measure in Medicare plans. All correlations were significant (p< 0.05).

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

**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

Same as below.

2014 Submission

The exclusions for this measure are based on clearly specified ICD-9-CM and ICD-10-CM codes for bilateral mastectomy. While these codes have not been tested in the context of this measure for validity, they are widely used across practitioners and considered to be valid. This measure does not allow for exclusions for patient refusal, provider refusal, or un-specified exclusions.

**2b2.2. What were the statistical results from testing exclusions**? (*include overall number and percentage of individuals excluded, frequency distribution of exclusions across measured entities, and impact on performance measure scores*)

N/A

**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*)

N/A

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

**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)*

2018 Submission

We assessed meaningful differences in performance of the measure in 2017 using the same methods specified below.

2014 Submission

To demonstrate meaningful differences in performance, NCQA calculates an interquartile range (IQR) for each indicator. The IQR is a measure of the dispersion of performance and is the difference between the 25th and 75th percentiles on a measure. To determine if the 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 percentiles. This 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 0.05, then the two plans’ performances are significantly different from each other. Using this method, we compared the performance rates of two randomly selected plans. We used these two plans as examples of 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*)

2018 Submission

HEDIS 2017 Variation in Performance Across Health Plans

Results are for the current measure assessing screening for women ages 50-74

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Avg. EP | Avg. | SD | 10th | 25th | 50th | 75th | 90th | IQR | p-value |
| Com. | 23,276 | 71 | 7 | 64 | 68 | 71 | 76 | 80 | 8 | <0.001 |
| Medicare | 7,944 | 72 | 10 | 61 | 67 | 73 | 79 | 83 | 12 | <0.001 |
| Medicaid | 4,769 | 59 | 9 | 48 | 53 | 59 | 66 | 70 | 13 | <0.001 |

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

IQR: Interquartile range

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

2014 Submission

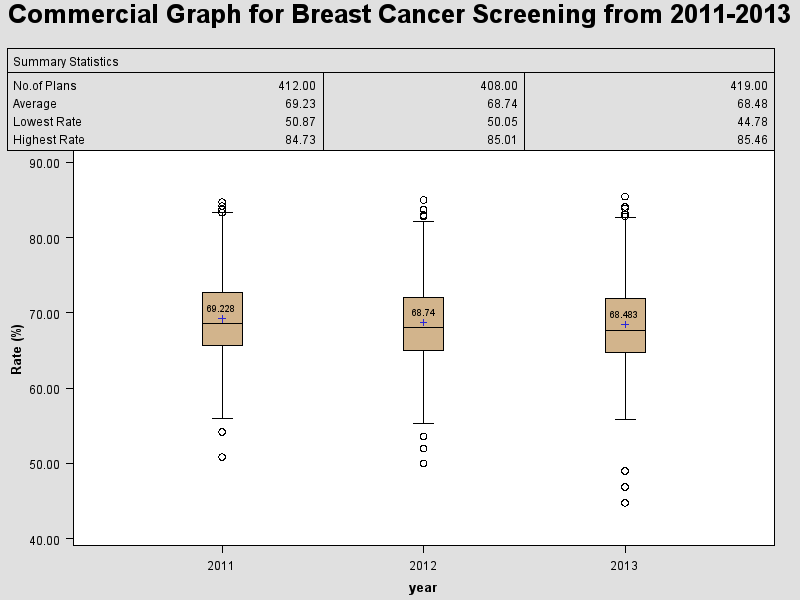
HEDIS 2013 Variation in Performance across Health Plans

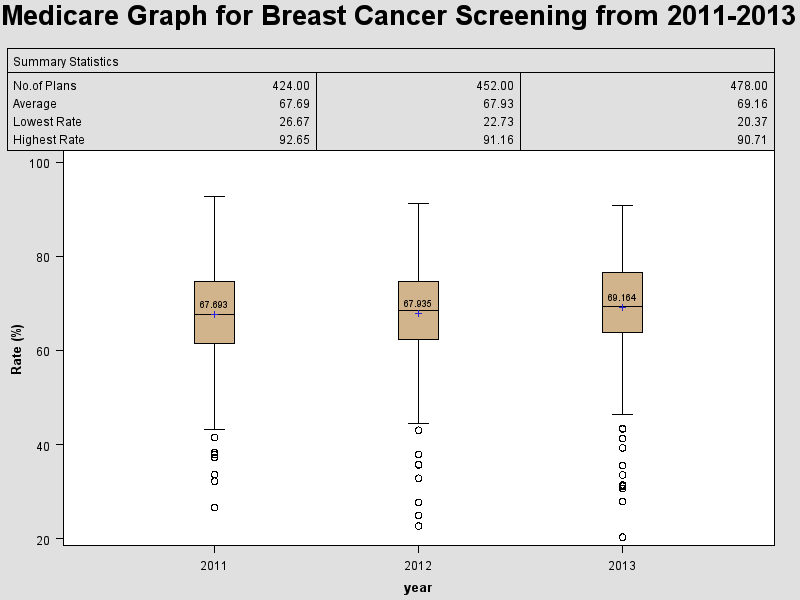
Note: results are from the measure specified for women 40-69 years, the most recent data available. The measure was updated in 2013 primarily to assess women 50-74 years.

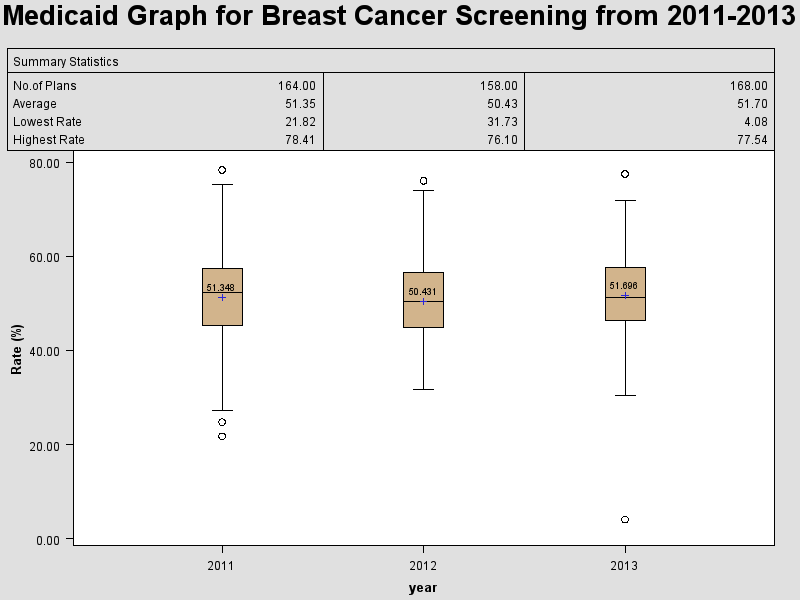
|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Avg. EP | Avg. | SD | 10th | 25th | 50th | 75th | 90th | Interquartile Range | p-value |
| Commercial HMO | 26,080 | 70.3 | 6.5 | 63.0 | 65.8 | 70.2 | 74.8 | 78.7 | 9.0 | <0.001 |
| Commercial PPO | 49,405 | 66.5 | 4.4 | 61.7 | 64.0 | 66.2 | 68.7 | 72.1 | 4.7 | <0.001 |
| Medicare HMO | 2,948 | 69.9 | 9.6 | 58.6 | 63.7 | 69.7 | 77.1 | 82.2 | 13.4 | <0.001 |
| Medicare PPO | 2,202 | 67.5 | 10.9 | 56.9 | 64.3 | 68.3 | 74.6 | 78.7 | 10.3 | <0.001 |
| Medicaid HMO | 4,065 | 51.9 | 9.1 | 41.7 | 46.5 | 51.5 | 57.9 | 62.9 | 11.4 | <0.001 |

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

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?*)

2018 Submission

The difference between the 25th and 75th percentile is statistically significant for all three product lines. For commercial plans, there is an 8 percentage point gap between 25th and 75th percentile plans. This gap represents an average 1,862 more patients that have been screened for breast cancer compared to low performing plans (estimated from average health plan eligible population).

2014 Submission

Average performance was 70% for Commercial and Medicare plans and 50% for Medicaid plans, with 10th percentile rates under 65%. The results show a 4-14% gap in performance between the 25th and 75th percentile-performing plans, which was statistically significant for all product lines and rates. Medicare HMOs had the largest performance gap with a 13.4 percentage point gap between the 25th and 75th percentiles. This gap represents on average 395 more patients receiving screening in high performing Medicare HMOs compared to low performing ones. All results suggest opportunities for improvement.

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

**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*)

2018 Submission

Same as below.

2014 Submission

This measure is collected with a complete sample; there are 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*)

2018 Submission

Same as below.

2014 Submission

This measure is collected with a complete sample; there are 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*)

2018 Submission

Same as below.

2014 Submission

This measure is collected with a complete sample; there are no missing data on this measure.