MEASURE WORKSHEET

This document summarizes the evaluation of the measure as it progresses through NQF's Consensus Development Process (CDP). The information submitted by measure developers/stewards is included after the Brief Measure Information, Preliminary Analysis, and Pre-meeting Public and Member Comments sections.

To navigate the links in the worksheet: Ctrl + click link to go to the link; ALT + LEFT ARROW to return

Brief Measure Information

NQF #: 2372
Measure Title: Breast Cancer Screening
Measure Steward: National Committee for Quality Assurance
Brief Description of Measure: Percentage of women 50-74 years of age who had a mammogram to screen for breast cancer
Developer Rationale: This measure assesses screening for breast cancer using mammography, which can prevent or detect early breast cancer, as well as reduce deaths from breast cancer. Early detection of breast cancer by mammography may also allow for a greater range of treatment options, including less-aggressive surgery and less-invasive therapy.

Numerator Statement: Women who received a mammogram to screen for breast cancer.
Denominator Statement: Women 50-74 years of age.
Denominator Exclusions: This measure excludes women with a history of bilateral mastectomy. The measure also excludes patients who use hospice services or are enrolled in an institutional special needs plan or living long-term in an institution any time during the measurement year.

Measure Type: Process
Data Source: Claims, Electronic Health Data
Level of Analysis: Health Plan, Integrated Delivery System

Original Endorsement Date: Sep 18, 2014 **Most Recent Endorsement Date:** Sep 18, 2014

Maintenance of Endorsement - Preliminary Analysis

To maintain NQF endorsement endorsed measures are evaluated periodically to ensure that the measures still meets the NQF endorsement criteria ("maintenance"). The emphasis for maintaining endorsement is focused on how effective the measure is for promoting improvements in quality. Endorsed measures should have some experience from the field to inform the evaluation. The emphasis for maintaining endorsement is noted for each criterion.

Criteria 1: Importance to Measure and Report

1a. Evidence
Maintenance measures – less emphasis on evidence unless there is new information or change in evidence since the prior evaluation.

1a. Evidence. The evidence requirements for a *structure, process or intermediate outcome* measure is that it is based on a systematic review (SR) and grading of the body of empirical evidence where the specific focus of the evidence matches what is being measured. For measures derived from patient report, evidence also should demonstrate that the target population values the measured process or structure and finds it meaningful.

The developer provides the following evidence for this measure:

- **Systematic Review of the evidence specific to this measure?** ☒ **Yes** ☐ **No**
- **Quality, Quantity and Consistency of evidence provided?** ☒ **Yes** ☐ **No**
- **Evidence graded?** ☒ **Yes** ☐ **No**

Evidence Summary

- The developer provides a logic model demonstrating the relationship between the focus of the measure and improved length and/or quality of life.
- The developer cites a 2016 United States Preventative Services Task Force (USPSTF) recommendation to support the measure as specified. The recommendation includes biennial screening mammography for women aged 50 to 74 years and received a B grade.
    - The measure numerator includes all of the following methods of mammograms: screening, diagnostic, film, digital or digital breast tomosynthesis (DBT). The USPSTF recommendation concludes that the current evidence is insufficient to assess the balance of benefits and harms of DBT as a primary screening method for breast cancer. This recommendation received an I grade.
- The developer cites a systematic review from Nelson et al. (2016) that includes more than 65 studies in support of the measure's focus, including 8 randomized control trials (RCTs).

Changes to evidence from last review

☐ **The developer attests that there have been no changes in the evidence since the measure was last evaluated.**

☒ **The developer provided updated evidence for this measure:**

Updates:

- The developer updated its submission to include the 2016 USPSTF recommendation; previous submissions included the 2009 recommendation. The focus and grade of the recommendation is unchanged.

Question for the Committee:

- *The evidence provided by the developer is updated, but is directionally the same compared to that for the previous NQF review. Does the Committee agree there is no need for repeat discussion and vote on Evidence?*

Guidance from the Evidence Algorithm

Box 1: no → Box 3: yes → Box 4: yes → Box 5b: yes → Moderate

Preliminary rating for evidence: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

1b. Gap in Care/Opportunity for Improvement and 1b. Disparities
Maintenance measures – increased emphasis on gap and variation

1b. Performance Gap. The performance gap requirements include demonstrating quality problems and opportunity for improvement.

The developer provides the following performance rates from HEDIS, which reflects the most recent years of measurement:

Commercial Plans (HMO and PPO combined)			
Measurement Year	**2015**	**2016**	**2017**
Mean	71.9%	71.4%	71.4%
Std. dev.	5.9%	6.5%	7.0%
Minimum	52.0%	37.2%	13.2%
10th percentile	64.8%	64.5%	64.0%
25th percentile	68.1%	67.7%	68.0%
50th percentile	71.1%	71.0%	71.2%
75th percentile	75.8%	75.4%	75.7%
90th percentile	80.5%	80.2%	79.8%
Maximum	89.6%	88.2%	88.0%

Medicaid Rates (HMO and PPO combined)

Measurement Year	**2015**	**2016**	**2017**
Mean	58.77%	58.51%	58.87%
Std. dev.	9.75%	10.09%	9.24%
Minimum	36.7%	17.78%	30.56%
10th percentile	45.99%	47.38%	48.0%
25th percentile	51.59%	52.28%	52.71%
50th percentile	58.37%	58.10%	59.02%
75th percentile	66.02%	65.06%	65.51%
90th percentile	71.32%	71.44%	70.29%
Maximum	87.88%	88.51%	87.92%

Medicare Rates (HMO and PPO combined)			
Measurement Year	**2015**	**2016**	**2017**
Mean	71.0%	72.4%	72.2%
Std. dev.	11.3%	09.8%	09.6%
Minimum	09.8%	14.3%	18.4%
10th percentile	60.5%	62.0%	61.4%
25th percentile	66.8%	67.4%	67.0%
50th percentile	71.7%	72.7%	73.0%
75th percentile	78.4%	79.2%	78.8%
90th percentile	82.9%	83.3%	82.9%
Maximum	92.1%	91.9%	91.1%

- From 2015 to 2017, performance rates for this measure have generally remained stable, with a decrease in performance in commercial plans, an increase in Medicare, and stable in Medicaid.
- The developer provided the following data for the denominator for the performance data for 2015, 2016, and 2017.

Commercial			
Measurement year	**2015**	**2016**	**2017**
Number of plans	410	426	420
Median denominator size by plan	7740	7510	7552

Medicaid			
Measurement year	**2015**	**2016**	**2017**
Number of plans	172	202	258
Median denominator size by plan	1642	1419	2439

Medicare			
Measurement year	**2015**	**2016**	**2017**
Number of plans	425	405	431

Median denominator size by plan	2173	2297	1890

Disparities

- HEDIS data are stratified by type of insurance (e.g., Commercial, Medicaid, Medicare). While not specified in the measure, this measure can also be stratified by demographic variables, such as race/ethnicity or socioeconomic status, in order to assess the presence of healthcare disparities, if the data are available to a plan.
- The developer provided disparities data from the literature, as follows:
    - One study found that mammography use in 2006 was 65 percent among white women and 59 percent among black women (Njai et al 2011).
    - African American women are more likely than white women to have longer intervals between screening mammograms, which may lead to an increase in later-stage cancer diagnoses (CDC 2012).
    - Between 2010 and 2014, breast cancer mortality for African American women was 41 percent higher than white women (Richardson et al 2016); one potential contributing factor to this health disparity is access to mammography screening services (Rust et al 2015).
    - National survey data also show that women who have attained lower degrees of education, lack health insurance coverage, or have lower socioeconomic status are less likely to have recently had a mammogram (National Center for Health Statistics 2015).

National Center for Health Statistics. 2015."Health, United States, 2015: With Special Feature on Racial and Ethnic Health Disparities". http://www.cdc.gov/nchs/data/hus/hus15.pdf#070 (Accessed Mar 13, 2018).

Njai, R., Siegel, P. Z., Miller, J. W., & Liao, Y. 2011. "Misclassification of Survey Responses and Black-White Disparity in Mammography Use, Behavioral Risk Factor Surveillance System, 1995-2006." Preventing Chronic Disease 8(3), A59.

Richardson, L.C., Henley, S.J., Miller, J.W., Massetti, G., and Thomas, C.C. 2016. "Patterns and Trends in Age-Specific Black-White Differences in Breast Cancer Incidence and Mortality – United States, 1999–2014." Morbidity and Mortality Weekly Report (MMWR) 65(40); 1093-1098. (November 30, 2016) http://www.cdc.gov/mmwr/volumes/65/wr/mm6540a1.htm.

Rust, G., Zhang, S., Malhotra, K., Reese, L., McRoy, L., Baltrus, P., Caplan, L. and Levine, R. 2015. "Paths to Health Equity – Local Area Variation in Progress Toward Eliminating Breast Cancer Mortality Disparities, 1990–2009." Cancer 121(16): 2765-2774. (December 1, 2016) doi: 10.1002/cncr.29405.

Questions for the Committee:

- *Is there a gap in care that warrants a national performance measure?*
- *Does this measure adequately address disparities?*

Preliminary rating for opportunity for improvement: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 1: Importance to Measure and Report (including 1a, 1b, 1c)

1a. Evidence

- The only new information that I am aware of is the issue of starting screening at age 45 instead of age 50. Should the committee discuss this change?
- Process measure- updated literature provided with similar results as shown in prior periods.
- 2016 USPSTF recommendation to support biennial screening mammography for women aged 50-74 years. B grade. Also systemic review and studies support the measure. The rating is moderate. One change to the measure is that digital breast tomosynthesis was added as an acceptable breast cancer screening method. However the evidence from USPSTF is listed as insufficient. The NCCN and ACR recommend the use of DBT for

primary screening of breast cancer. I don't know if additional discussion of the evidence would be productive. The measure is adjusted to make sure that it is accounting for evolving clinical practice. The measure isn't dictating practice, but reacting to changes in recommendations and practice.

- I agree with the staff conclusions. I note, however, that the 2016 USPSTF report is based on 7 (not 8) RCTs, and they are all "fair" quality. I do not see a reason to discuss and vote on this again.
- Strong evidence
- Process measure, based on logic model and systematic review, updated in 2016, that led to a US Preventive Services Task Force recommendation of B for mammography for women aged 50-74. Evidence grades at least moderate.
- Process measure evaluating breast cancer screening by mammogram for women 50-74 years of age. Empirical evidence provided with systematic review, QQC, and grading of evidence. Evidence updated with 2016 USPSTF recommendation since measure last evaluated. Per evidence algorithm, evidence rating is moderate.

1b. Performance Gap

- The performance gap is sufficient to warrant the continued use of this measure. There are significant disparities in population subgroups.
- Performance rates relatively unchanged across payers- decrease in commercial plans, increase in Medicare and same status for Medicaid. Disparity noted with African-American women as increase in cancer rates later in age potentially related to reduced screening mammography in line with this measure.
- The developers share results that show a disparity in screening based on plan type (Commercial, Medicaid and Medicare). The literature and and analysis of state-level survey and clinical data have shown racial/ethnic and socio-economic disparities in screening, stage at diagnosis, care and mortality. The gap warrants a national performance measure. The measure itself is neutral, I'm not sure how a measure can address disparities. However, if accurate demographic data are collected additional analyses based on these demographics can be done to determine the gaps in screening. This measure along with others could be used together to look at disparities in stage at diagnosis, access to care, treatment and mortality.
- There is definitely a gap, as these data clearly show. The measure submission speaks about stratifying by the proportion of a plans's membership that is minority, and this is not the same as stratifying on individual characteristics. But this doesn't change my conclusion that there are important performance gaps.
- Persistent gap remains
- HEDIS data demonstrate a gap in performance, with a range from 25%ile to 75%ile of 68-76% in commercial plans and 52-65% in Medicaid plans and 67-79% in the Medicare plans, for 2017. Racial disparities have been documented in 4 literature studies.
- Performance gap still exists, particularly in Medicaid population. Slight improvement over last 3 years. Racial/ethnic/socio-economic/age, and geographic disparities exist. Moderate to high performance gap.

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability: Specifications and Testing
2b. Validity: Testing; Exclusions; Risk-Adjustment; Meaningful Differences; Comparability Missing Data
2c. For composite measures: empirical analysis support composite approach

Reliability

2a1. Specifications requires the measure, as specified, to produce consistent (reliable) and credible (valid) results about the quality of care when implemented. For maintenance measures – no change in emphasis – specifications should be evaluated the same as with new measures.

2a2. Reliability testing demonstrates if 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 enough to distinguish differences in performance across providers. For maintenance measures – less emphasis if no new testing data provided.

Validity

2b2. Validity testing should demonstrate the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality. For maintenance measures – less emphasis if no new testing data provided.

2b2-2b6. Potential threats to validity should be assessed/addressed.

Composite measures only:

2d. Empirical analysis to support composite construction. Empirical analysis should demonstrate that the component measures add value to the composite and that the aggregation and weighting rules are consistent with the quality construct.

Complex measure evaluated by Scientific Methods Panel? ☐ **Yes** ☒ **No**
Evaluators: Staff—Staff Analysis

Questions for the Committee regarding reliability:

- *Do you have any concerns that the measure can be consistently implemented (i.e., are measure specifications adequate)?*
- *The staff is satisfied with the reliability testing for the measure; the signal-to-noise ratio is considered very high, per the developer. Does the Committee think there is a need to discuss and/or vote on Reliability?*

Questions for the Committee regarding validity:

- *Do you have any concerns regarding the validity of the measure (e.g., exclusions, risk-adjustment approach, etc.)? The developer states it did not perform an analysis on the effect of the specified exclusions, but notes they are specified on ICD-9 and ICD-10 codes that are widely used and considered to be valid.*
- *The construct validity testing hypothesizes a relationship between colorectal cancer screening and the breast cancer screening, and the developer reports a strong, positive relationship. Does the Committee think there is a need to discuss or vote on Validity?*
-

Preliminary rating for reliability: ☒ **High** ☐ **Moderate** ☐ **Low** ☐ **Insufficient**

Preliminary rating for validity: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 2: Scientific Acceptability of Measure Properties (including all 2a, 2b, and 2c)

2a1. Reliability – Specifications

- All data elements and steps are clear. I have no concerns.
- Data taken from claims history- deemed reliable with little potential for discrepancy among payer reporting. No concerns noted for reliability.
- Measure is in use.
- No concerns, and no reason to discuss.
- Ok
- Measure specifications are good, and elements clearly defined. No concerns.
- Measure specifications are clearly defined. Exclusions are clearly defined. No risk adjustment done. Measure is likely to be implemented consistently.

2a2. Reliability -- (Testing)

- No
- No threats to reliability noted.
- I don't think that there is a need to discuss reliability and validity of this measure.
- No concerns, and no reason to discuss.
- No
- Reliability is excellent, no concerns. Based on >1000 health plans and >1500 patients/plan data from 2017 to HEDIS.
- Reliability testing was done using the beta-binomial method to test the "signal-to-noise" of the measure using HEDIS data elements from commercial, Medicaid, and Medicare plans. Overall reliability was 0.95 and higher signifying high reliability.

2b1. Validity -Testing
2b4-7. Threats to Validity (Statistically Significant Differences, Multiple Data Sources, Missing Data)
2b4. Meaningful Differences

- 2b1. no concerns; 2b4.-7. There are no threats to validity

- Two concerns for validity reported- 1. Exclusions based upon ICD9-ICD10 codes for bilateral mastectomy not tested for mammography but widely used and considered valid and 2. Does not allow for exclusions of patient refusal, provider refusal or unspecified exclusions.
- No Missing
- No concerns, and no reason to discuss. It is possible that a woman gets a mammogram outside the plan, and that this would be "missing," but that is not likely to be a problem.
- Testing whether "the difference between the 25th and 75th percentile is statistically significant" is an odd, and probably invalid, way to assess validity, but other evidence is enough to establish validity.
- Not a threat to validity.
- Empiric (construct) validity testing was done using a Pearson correlation test. Results demonstrated 0.71 correlation in commercial plans and 0.72 correlation in Medicare plans. Face validity was not with previous submission with "good agreement." Overall moderate validity.

2b2-3. Other Threats to Validity (Exclusions, Risk Adjustment)

- 2b2. No to first question. No to second question. 2b3. N/A
- Not risk adjusted- not applicable.
- Exclusions are consistent, but women living in institutional SNP or living long-term in an institution at any time during the measurement year are excluded. I'm not sure of the rationale for these exclusions. What does SNP mean? Also those who refused aren't excluded and should be, but I doubt that there is a way to capture this information.
- NA
- No risk adjustment. exclusions appropriate
- N/A
- Several exclusions noted for measure, but no exclusion analysis done. No risk adjustment done.

Criterion 3. Feasibility
Maintenance measures – no change in emphasis – implementation issues may be more prominent

3. Feasibility is the extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

- This measure is specified for administrative claims data. All data elements are in defined fields in electronic claims.
- Data are generated or collected by and used by healthcare personnel during the provision of care (e.g., blood pressure, lab value, diagnosis, depression score), coded by someone other than person obtaining original information (e.g., DRG, ICD-9 codes on claims), abstracted from a record by someone other than person obtaining original information (e.g., chart abstraction for quality measure or registry).
- Commercial use of the measure requires prior written consent from NCQA. Noncommercial uses do not require consent. Use by physicians in connection with their own practices is not considered commercial use.

Question for the Committee:

- *Are the required data elements routinely generated and used during care delivery?*

Preliminary rating for feasibility: ☒ **High** ☐ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 3: Feasibility

3. Feasibility

- All data elements are routinely generated and available in electronic form. I have no concerns about putting this measure into operational use.
- No concerns re: feasibility, agree with high rating.
- Data elements are routinely generated and collected.
- No concerns.

- Long track record of feasibility
- Feasibility is high.
- Data collection obtained through administrative data, and all data elements are in defined, electronic claims fields. High feasibility.

Criterion 4: Usability and Use
Maintenance measures – increased emphasis – much greater focus on measure use and usefulness, including both impact/improvement and unintended consequences

4a. Use (4a1. Accountability and Transparency; 4a2. Feedback on measure)

4a. Use evaluate the extent to which audiences (e.g., consumers, purchasers, providers, policymakers) use or could use performance results for both accountability and performance improvement activities.

4a.1. Accountability and Transparency. Performance results are used in at least one accountability application within three years after initial endorsement and are publicly reported within six years after initial endorsement (or the data on performance results are available). If not in use at the time of initial endorsement, then a credible plan for implementation within the specified timeframes is provided.

Current uses of the measure
Publicly reported? ☒ **Yes** ☐ **No**

Current use in an accountability program? ☒ **Yes** ☐ **No** ☐ **UNCLEAR**
OR
Planned use in an accountability program? ☐ **Yes** ☒ **No**

Accountability program details

- This measure is used in many public reporting and payment programs, including: CMS Medicare Star Rating Program, CMS Medicaid Adult Core Set, CMS Quality Payment Program, California's Value based Pay for Performance Program, and CMS Qualified Health Plan (QHP) Quality Rating System (QRS).
- This measure also is used in two quality improvement programs: NCQA Quality Compass and NCQA State of Health Care Quality.

4a.2. Feedback on the measure by those being measured or others. Three criteria demonstrate feedback: 1) those being measured have been given performance results or data, as well as assistance with interpreting the measure results and data; 2) those being measured and other users have been given an opportunity to provide feedback on the measure performance or implementation; 3) this feedback has been considered when changes are incorporated into the measure

Feedback on the measure by those being measured or others

- Questions received through the Policy Clarification Support system have generally sought clarification about the mammography screening methods that satisfy the measure numerator.
- During the measure's last major update, feedback informed how the developer revised the measure to include digital breast tomosynthesis as a new screening method.

***Questions for the Committee**:*

- *How have (or can) the performance results be used to further the goal of high-quality, efficient healthcare?*
- *How has the measure been vetted in real-world settings by those being measured or others?*

Preliminary rating for Use: ☒ **Pass** ☐ **No Pass**

4b. Usability (4a1. Improvement; 4a2. Benefits of measure)

4b. Usability evaluate the extent to which audiences (e.g., consumers, purchasers, providers, policymakers) use or could use performance results for both accountability and performance improvement activities.

4b.1 Improvement. Progress toward achieving the goal of high-quality, efficient healthcare for individuals or populations is demonstrated.

Improvement results

- From 2011 to 2017 average performance rates increased by about 6% for Medicare and Medicaid plans and were steady in commercial plans.
- Variation between the 10th and 90th percentile suggests room for improvement.

4b2. Benefits vs. harms. Benefits of the performance measure in facilitating progress toward achieving high-quality, efficient healthcare for individuals or populations outweigh evidence of unintended negative consequences to individuals or populations (if such evidence exists).

Unexpected findings (positive or negative) during implementation

- None reported by the developer.

Potential harms

- The developer identified one potential unintended consequence of the Breast Cancer Screening measure: too-frequent screening. Feedback from an advisory panel indicated that, in an effort to meet the two-year requirement, women often are encouraged to seek screening earlier than the two-year mark. In order to address potential over-screening, NCQA adjusted the numerator time frame to 27 months, providing a three-month leeway to account for the logistics of scheduling and receiving a mammogram.

***Questions for the Committee**:*

- *How can the performance results be used to further the goal of high-quality, efficient healthcare?*
- *Do the benefits of the measure outweigh any potential unintended consequences?*

Preliminary rating for Usability and use: ☒ **High** ☐ **Moderate** ☐ **Low** ☐ **Insufficient**
RATIONALE:

Committee pre-evaluation comments

Criteria 4: Usability and Use

4a. Use

- 4a1. This measure is routinely being reported to health plans and publicly. 4a2. Yes to all questions.
- Used widely in accountability programs;
- This is one of the most widely used measures for both accountability and QI, and I see no problems.
- Widely used
- NCQA uses for public reporting, quality improvement and gives feedback. CMS uses for accountability programs.

4b. Usability

- 4b1. Continue to report results back to health plans and providers. 4b2. Unintended consequences well described and discussed by measure stewards.
- Only concern raised was frequency may be sooner than 2 years in order to meet the timeline goal;
- One of the unintended consequences mentioned is too frequent screening. The numerator time frame was broadened to 27 months to address this. I'm also wondering if including digital breast tomosynthesis would encourage its use as primary

screening method, when the USPSTF concluded that the current evidence is insufficient assess the benefits and harms? On the other hand if you don't include DBT in the numerator, you will exclude women who use it and underestimate the numerator.
- No concerns.
- Need new strategies to close performance gap. Would be even better if this were truly a population-based measure
- Increasing performance rates during past 7 years suggests improvement in quality care. Unintended consequences identified are in the area of over diagnosis and overtreatment of non-malignant breast disease, so the benefits of identifying cancer are outweighed in these age groups as identified by the USPSTF and its B recommendation.
- Measure in current use in many public reporting programs (NCQA, CMS, California Value Based Pay for Performance Program) and other payment programs (CMS MA Stars Rating, CMS QPP, California Value Based Pay for Performance Program), High usability and use.

Criterion 5: Related and Competing Measures

Related or competing measures

- 0508 : Diagnostic Imaging: Inappropriate Use of "Probably Benign" Assessment Category in Screening Mammograms (American College of Radiology)
- 0509 : Diagnostic Imaging: Reminder System for Screening Mammograms (American College of Radiology)

Harmonization

- These two measures are stewarded by the American College of Radiology and are specified at the clinician level rather than the health plan level; they are related, not competing.
- The developer states that the measures are harmonized to the extent possible.

Committee pre-evaluation comments

Criterion 5: Related and Competing Measures

N/A

Public and member comments

Comments and Member Support/Non-Support Submitted as of: Month/Day/Year

- **Of the XXX NQF members who have submitted a support/non-support choice:**
  - XX support the measure
  - YY do not support the measure

Measure Number: 2372
Measure Title: Breast Cancer Screening

Scientific Acceptability: Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. Measures must be judged to meet the subcriteria for both reliability and validity to pass this criterion

Instructions for filling out this form:

- Please complete this form for each measure you are evaluating.
- Please pay close attention to the skip logic directions. ***Directives that require you to skip questions are marked in red font.***
- If you are unable to check a box, please highlight or shade the box for your response.
- You must answer the "overall rating" item for both Reliability and Validity. Also, be sure to answer the composite measure question at the end of the form if your measure is a composite.
- For several questions, we have noted which sections of the submission documents you should ***REFERENCE*** and provided ***TIPS*** to help you answer them.
- ***It is critical that you explain your thinking/rationale if you check boxes that require an explanation.*** Please add your explanation directly below the checkbox in a different font color. Also, feel free to add additional explanation, even if you select a checkbox where an explanation is not requested (if you do so, please type this text directly below the appropriate checkbox).
- ***Please refer to the Measure Evaluation Criteria and Guidance document (pages 18-24) and the 2-page Key Points document when evaluating your measures***. This evaluation form is an adaptation of Alogorithms 2 and 3, which provide guidance on rating the Reliability and Validity subcriteria.
- ***Remember*** that testing at either the data element level **OR** the measure score level is accepted for some types of measures, but not all (e.g., instrument-based measures, composite measures), and therefore, the embedded rating instructions may not be appropriate for all measures.
- ***Please base your evaluations solely on the submission materials provided by developers.*** NQF strongly discourages the use of outside articles or other resources, even if they are cited in the submission materials. If you require further information or clarification to conduct your evaluation, please communicate with NQF staff (methodspanel@qualityforum.org).

RELIABILITY

1. Are submitted specifications precise, unambiguous, and complete so that they can be consistently implemented?

   REFERENCE: "MIF_xxxx" document
   ***NOTE**: NQF staff will conduct a separate, more technical, check of eMeasure (eCQM) specifications, value sets, logic, and feasibility, so no need to consider these in your evaluation.*
   ***TIPS**: Consider the following: Are all the data elements clearly defined? Are all appropriate codes included? Is the logic or calculation algorithm clear? Is it likely this measure can be consistently implemented?*

   ☒Yes (go to Question #2)

   ☐No (please explain below, and go to Question #2) NOTE that even though ***non-precise specifications should result in an overall LOW rating for reliability***, we still want you to look at the testing results.

2. Was empirical reliability testing (at the data element or measure score level) conducted using statistical tests with the measure as specified?

   REFERENCE: "MIF_xxxx" document for specifications, testing attachment questions 1.1-1.4 and section 2a2
   ***TIPS**: Check the "NO" box below if: only descriptive statistics are provided; only describes process for data management/cleaning/computer programming; testing does not match measure specifications (i.e., data source, level of analysis, included patients, etc.)*

   ☒Yes (go to Question #3)

   ☐No, there is reliability testing information, but *not* using statistical tests and/or not for the measure as specified **OR** there is no reliability testing (please explain below, skip Questions #3-8, then go to Question #9)

3. Was reliability testing conducted with computed performance measure scores for each measured entity?

   REFERENCE: "Testing attachment_xxx", section 2a2.1 and 2a2.2
   ***TIPS**: Answer no if: only one overall score for all patients in sample used for testing patient-level data*

   ☒Yes (go to Question #4)

   ☐No (skip Questions #4-5 and go to Question #6)

4. Was the method described and appropriate for assessing the proportion of variability due to real differences among measured entities? *NOTE: If multiple methods used, at least one must be appropriate.*

   REFERENCE: Testing attachment, section 2a2.2
   ***TIPS**: Examples of appropriate methods include signal-to-noise analysis (e.g. Adams/RAND tutorial); random split-half correlation; other accepted method with description of how it assesses reliability of the performance score.*

   ☒Yes (go to Question #5)

   ☐No (please explain below, then go to question #5 and rate as INSUFFICIENT)

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 developer used a beta-binomial method to test the "signal-to-noise" of the measure, where the proportion of total variation attributable to a health plan is the signal, and the proportion attributable to measurement error is the noise. Reliability is represented as a ratio.

Below is a description of the sample used for measure score reliability testing. 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

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.

5. **RATING (score level)** - What is the level of certainty or confidence that the performance measure scores are reliable?

   REFERENCE: Testing attachment, section 2a2.2
   ***TIPS**: Consider the following: Is the test sample adequate to generalize for widespread implementation? Do the results demonstrate sufficient reliability so that differences in performance can be identified?*

☒High (go to Question #6)
☐Moderate (go to Question #6)
☐Low (please explain below then go to Question #6)
☐Insufficient (go to Question #6)

Overall reliability across all three plan types was 0.95 and higher, which the developer states indicating high reliability.

This table summarizes the variability of individual plan reliability.

	Overall Reliability (beta binomial)
Commercial (n=419)	0.998
Medicaid (n=168)	0.993
Medicare (n=478)	0.997

6. Was reliability testing conducted with patient-level data elements that are used to construct the performance measure?
   REFERENCE: Testing attachment, section 2a2.
   TIPS*: Prior reliability studies of the same data elements may be submitted; if comparing abstraction to "authoritative source/gold standard" go to Question #9)*
   ☐Yes (go to Question #7)
   ☒No (if there is score-level testing that you rated something other than INSUFFICIENT in Question #5, skip questions #7-9, then go to Question #10 (OVERALL RELIABILITY); otherwise, skip questions #7-8 and go to Question #9)

7. Was the method described and appropriate for assessing the reliability of ALL critical data elements?
   REFERENCE: Testing attachment, section 2a2.2
   TIPS*: For example: inter-abstractor agreement (ICC, Kappa); other accepted method with description of how it assesses reliability of the data elements*
   Answer no if: only assessed percent agreement; did not assess separately for all data elements (at least numerator, denominator, exclusions)
   ☐Yes (go to Question #8)
   ☐No (if no, please explain below, then go to Question #8 and rate as INSUFFICIENT)

8. **RATING (data element)** – Based on the reliability statistic and scope of testing (number and representativeness of patients and entities), what is the level of certainty or confidence that the data used in the measure are reliable?
   REFERENCE: Testing attachment, section 2a2
   TIPS*: Consider the following: Is the test sample adequate to generalize for widespread implementation? Can data elements be collected consistently?*
   ☐Moderate (skip Question #9 and go to Question #10 (OVERALL RELIABILITY); if score-level testing was NOT conducted, rate Question #10 as MODERATE)
   ☐Low (skip Question #9 and go to Question #10 (OVERALL RELIABILITY); if score-level testing was NOT conducted, rate Question #10 (OVERALL RELIABILITY) as LOW)
   ☐Insufficient (go to Question #9)

9. Was **empirical VALIDITY testing** of patient-level data conducted?
   REFERENCE: testing attachment section 2b1.
   NOTE: Skip this question if empirical reliability testing was conducted and you have rated Question #5 and/or #8 as anything other than INSUFFICIENT)

TIP: *You should answer this question ONLY if score-level or data element reliability testing was NOT conducted or if the methods used were NOT appropriate. For most measures, NQF will accept data element validity testing in lieu of reliability testing—but* ***check with NQF staff before proceeding, to verify.***

☐Yes (go to Question #10 and answer using your rating from data element validity testing – Question #23)

☐No (please explain below, go to Question #10 (OVERALL RELIABILITY) and rate it as INSUFFICIENT. Then go to Question #11.)

OVERALL RELIABILITY RATING

10. **OVERALL RATING OF RELIABILITY** taking into account precision of specifications (see Question #1) and all testing results:

☒High (NOTE: Can be HIGH only if score-level testing has been conducted)

☐Moderate (NOTE: Moderate is the highest eligible rating if score-level testing has not been conducted)

☐Low (please explain below) [NOTE: Should rate LOW if you believe specifications are NOT precise, unambiguous, and complete]

☐Insufficient (please explain below) [NOTE: For most measure types, testing at both the score level and the data element level is not required, but check with NQF staff]

(Box 1) yes → (Box 2) yes → (Box 4) yes → (Box 5) yes → (Box 6a) yes → High rating

VALIDITY

Assessment of Threats to Validity

11. Were potential threats to validity that are relevant to the measure empirically assessed ()?

REFERENCE: Testing attachment, section 2b2-2b6

TIPS*: Threats to validity that should be assessed include: exclusions; need for risk adjustment; ability to identify statistically significant and meaningful differences; multiple sets of specifications; missing data/nonresponse.*

☐Yes (go to Question #12)

☒No (please explain below and then go to Question #12) [NOTE that ***non-assessment of applicable threats should result in an overall INSUFFICENT rating for validity***]

The developer did not provide an analysis of measure exclusions.

12. Analysis of potential threats to validity: Any concerns with measure exclusions?

REFERENCE: Testing attachment, section 2b2.

TIPS*: Consider the following: Are the exclusions consistent with the evidence? Are any patients or patient groups inappropriately excluded from the measure? Are the exclusions/exceptions of sufficient frequency and variation across providers to be needed (and outweigh the data collection burden)? If patient preference (e.g., informed decisionmaking) is a basis for exclusion, does it impact performance and if yes, is the measure specified so that the information about patient preference and the effect on the measure is transparent?*

☒Yes (please explain below then go to Question #13)

☐No (go to Question #13)

☐Not applicable (i.e., there are no exclusions specified for the measure; go to Question #13)

The measure exclusions are based on specified ICD-9 and ICD-10 codes for bilateral mastectomy, which have not been tested in the context of this measures but are widely used and considered to be valid.

This measure does not allow for exclusions for patient refusal, provider refusal, on un-specified exclusions.

13. Analysis of potential threats to validity: Risk-adjustment (this applies to all outcome, cost, and resource use measures and "NOT APPLICABLE" is not an option for those measures; the risk-adjustment questions (13a-13c, below) also may apply to other types of measures)

REFERENCE: Testing attachment, section 2b3.

13a. Is a conceptual rationale for social risk factors included? ☐Yes ☐No

13b. Are social risk factors included in risk model? ☐Yes ☐No

13c. Any concerns regarding the risk-adjustment approach?

***TIPS**: Consider the following: **If measure is risk adjusted**: If the developer asserts there is **no conceptual basis** for adjusting this measure for social risk factors, do you agree with the rationale? Are the candidate and final variables included in the risk adjustment model adequately described for the measure to be implemented? Are the candidate and final variables included in the risk adjustment model adequately described for the measure to be implemented? Are all of the risk adjustment variables present at the start of care (if not, do you agree with the rationale)? If social risk factors are not included in the risk-adjustment approach, do you agree with the developer's decision? Is an appropriate risk-adjustment strategy included in the measure (e.g., adequate model discrimination and calibration)? Are all statistical model specifications included, including a "clinical model only" if social risk factors are included in the final model? If a measure is NOT risk-adjusted, is a justification for **not risk adjusting** provided (conceptual and/or empirical)? Is there any evidence that contradicts the developer's rationale and analysis for not risk-adjusting?*

☐Yes (please explain below then go to Question #14)

☐No (go to Question #14)

☒Not applicable (e.g., this is a structure or process measure that is not risk-adjusted; go to Question #14)

14. Analysis of potential threats to validity: Any concerns regarding ability to identify meaningful differences in performance or overall poor performance?

REFERENCE: Testing attachment, section 2b4.

☐Yes (please explain below then go to Question #15)

☒No (go to Question #15)

15. Analysis of potential threats to validity: Any concerns regarding comparability of results if multiple data sources or methods are specified?

REFERENCE: Testing attachment, section 2b5.

☐Yes (please explain below then go to Question #16)

☐No (go to Question #16)

☒Not applicable (go to Question #16)

16. Analysis of potential threats to validity: Any concerns regarding missing data?

REFERENCE: Testing attachment, section 2b6.

☒Yes (please explain below then go to Question #17)

☐No (go to Question #17)

Assessment of Measure Testing

17. Was <u>empirical</u> validity testing conducted using the measure as specified and with appropriate statistical tests?

REFERENCE: Testing attachment, section 2b1.

***TIPS**: Answer no if: only face validity testing was performed; only refer to clinical evidence; only descriptive statistics; only describe process for data management/cleaning/computer programming; testing does not match measure specifications (i.e. data, eMeasure, level, setting, patients).*

☒Yes (go to Question #18)

☐No (please explain below, then skip Questions #18-23 and go to Question #24)

18. Was validity testing conducted with <u>computed performance measure scores</u> for each measured entity?

REFERENCE: Testing attachment, section 2b1.

***TIPS**: Answer no if: one overall score for all patients in sample used for testing patient-level data.*

☒Yes (go to Question #19)

☐No (please explain below, then skip questions #19-20 and go to Question #21)

19. Was the method described and appropriate for assessing conceptually and theoretically sound hypothesized relationships?

REFERENCE: Testing attachment, section 2b1.

***TIPS**: For example: correlation of the performance measure score on this measure and other performance measures; differences in performance scores between groups known to differ on quality; other accepted method with description of how it assesses validity of the performance score*

☒Yes (go to Question #20)

☐No (please explain below, then go to Question #20 and rate as INSUFFICIENT)

The developer conducted construct validity testing to determine the correlation between this measure and measures that are hypothesized to be related. The developer used a Pearson Correlation test to examine the association between this measure and NQF #0034 Colorectal Cancer Screening and NQF #0032 Cervical Cancer Screening.

20. **RATING (measure score)** - Based on the measure score results (significance, strength) and scope of testing (number of measured entities and representativeness) and analysis of potential threats, what is the level of certainty or confidence that the performance measure scores are a valid indicator of quality?

☐High (go to Question #21)

☒Moderate (go to Question #21)

☐Low (please explain below then go to Question #21)

☐Insufficient (go to Question #21)

The measure had a strong positive correlation to the measure Colorectal Cancer Screening in both Commercial plans (0.71) and in Medicare plans (0.72).

21. Was validity testing conducted with patient-level data elements?

REFERENCE: Testing attachment, section 2b1.

***TIPS**: Prior validity studies of the same data elements may be submitted*

☐Yes (go to Question #22)

☒No (if there is score-level testing that you rated something other than INSUFFICIENT in Question #20, skip questions #22-25, and go to Question #26 (OVERALL VALIDITY); otherwise, skip questions #22-23 and go to Question #24)

22. Was the method described and appropriate for assessing the accuracy of ALL critical data elements? *NOTE that data element validation from the literature is acceptable.*

***REFERENCE:** Testing attachment, section 2b1.*

***TIPS**: For example: Data validity/accuracy as compared to authoritative source- sensitivity, specificity, PPV, NPV; other accepted method with description of how it assesses validity of the data elements.*

Answer No if: only assessed percent agreement; did not assess separately for all data elements (at least numerator, denominator, exclusions)

☐Yes (go to Question #23)

☐No (please explain below, then go to Question #23 and rate as INSUFFICIENT)

23. **RATING (data element)** - Based on the data element testing results (significance, strength) and scope of testing (number and representativeness of patients and entities) and analysis of potential threats, what is the level of certainty or confidence that the data used in the measure are valid?

☐Moderate (skip Questions #24-25 and go to Question #26)

☐Low (please explain below, skip Questions #24-25 and go to Question #26)

☐Insufficient (go to Question #24 only if no other empirical validation was conducted OR if the measure has not been previously endorsed; otherwise, skip Questions #24-25 and go to Question #26)

24. Was face validity systematically assessed by recognized experts to determine agreement on whether the computed performance measure score from the measure as specified can be used to distinguish good and poor quality?

NOTE: If appropriate empirical testing has been conducted, then evaluation of face validity is not necessary; you should skip this question and Question 25, and answer Question #26 based on your answers to Questions #20 and/or #23]

REFERENCE: Testing attachment, section 2b1.

TIPS: *Answer no if: focused on data element accuracy/availability/feasibility/other topics; the degree of consensus and any areas of disagreement not provided/discussed.*

☒Yes (go to Question #25)

☐No (please explain below, skip question #25, go to Question #26 (OVERALL VALIDITY) and rate as INSUFFICIENT)

The face validity of the measure was last assessed in 2012-2013. The developer did not provide the results of the face validity assessment, but noted that the assessment concluded with good agreement that the measure as specified accurately assess breast cancer screening in health plans.

25. **RATING (face validity)** - Do the face validity testing results indicate substantial agreement that the performance measure score from the measure as specified can be used to distinguish quality AND potential threats to validity are not a problem, OR are adequately addressed so results are not biased?

REFERENCE: Testing attachment, section 2b1.

TIPS: *Face validity is no longer accepted for maintenance measures unless there is justification for why empirical validation is not possible and you agree with that justification.*

☒Yes (if a NEW measure, go to Question #26 (OVERALL VALIDITY) and rate as MODERATE)

☒ Yes (if a MAINTENANCE measure, do you agree with the justification for not conducting empirical testing? If no, go to Question #26 (OVERALL VALIDITY) and rate as INSUFFICIENT; otherwise, rate Question #26 as MODERATE)

☐No (please explain below, go to Question #26 (OVERALL VALIDITY) and rate AS LOW)

OVERALL VALIDITY RATING

26. **OVERALL RATING OF VALIDITY** taking into account the results and scope of all testing and analysis of potential threats.

☐High (NOTE: Can be HIGH only if score-level testing has been conducted)

☒Moderate (NOTE: Moderate is the highest eligible rating if score-level testing has NOT been conducted)

☐Low (please explain below) [NOTE: Should rate LOW if you believe that there are threats to validity and/or threats to validity were not assessed]

☐Insufficient (if insufficient, please explain below) [NOTE: For instrument-based measures and some composite measures, testing at both the score level and the data element level is required; if not conducted, should rate as INSUFFICIENT—please check with NQF staff if you have questions.]

(Box 1) yes → (Box 2) yes → (Box 5) yes → (Box 6) yes → (Box 7b) yes → Moderate rating

NATIONAL QUALITY FORUM—Evidence (subcriterion 1a)

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

Measure Title: Breast Cancer Screening

IF the measure is a component in a composite performance measure, provide the title of the Composite Measure here: Click here to enter composite measure #/ title

Date of Submission: 4/16/2018

Instructions

- *Complete 1a.1 and 1a.2 for all measures. If instrument-based measure, complete 1a.3.*
- *Complete **EITHER 1a.2, 1a.3 or 1a.4** as applicable for the type of measure and evidence.*
- *For composite performance measures:*
  - *A separate evidence form is required for each component measure unless several components were studied together.*
  - *If a component measure is submitted as an individual performance measure, attach the evidence form to the individual measure submission.*
- All information needed to demonstrate meeting the evidence subcriterion (1a) must be in this form. An appendix of *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.
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.

Note: The information provided in this form is intended to aid the Standing Committee and other stakeholders in understanding to what degree the evidence for this measure meets NQF's evaluation criteria.

1a. Evidence to Support the Measure Focus

The measure focus is evidence-based, demonstrated as follows:

- Outcome: [3] Empirical data demonstrate a relationship between the outcome and at least one healthcare structure, process, intervention, or service. If not available, wide variation in performance can be used as evidence, assuming the data are from a robust number of providers and results are not subject to systematic bias.
- Intermediate clinical outcome: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured intermediate clinical outcome leads to a desired health outcome.
- Process: [5] a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured process leads to a desired health outcome.
- Structure: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured structure leads to a desired health outcome.
- Efficiency: [6] evidence not required for the resource use component.
- For measures derived from patient reports, evidence should demonstrate that the target population values the measured outcome, process, or structure and finds it meaningful.
- Process measures incorporating Appropriate Use Criteria: See NQF's guidance for evidence for measures, in general; guidance for measures specifically based on clinical practice guidelines apply as well.

Notes

3. Generally, rare event outcomes do not provide adequate information for improvement or discrimination; however, serious reportable events that are compared to zero are appropriate outcomes for public reporting and quality improvement.

4. The preferred systems for grading the evidence are the Grading of Recommendations, Assessment, Development and Evaluation (GRADE) guidelines and/or modified GRADE.

5. Clinical care processes typically include multiple steps: assess → identify problem/potential problem → choose/plan intervention (with patient input) → provide intervention → evaluate impact on health status. If the measure focus is one step in such a multistep process, the step with the

strongest evidence for the link to the desired outcome should be selected as the focus of measurement. Note: A measure focused only on collecting PROM data is not a PRO-PM.

6. Measures of efficiency combine the concepts of resource use and quality (see NQF's Measurement Framework: Evaluating Efficiency Across Episodes of Care; AQA Principles of Efficiency Measures).

1a.1.This is a measure of: (*should be consistent with type of measure entered in De.1*)

Outcome

☐ Outcome: Click here to name the health outcome

☐Patient-reported outcome (PRO): Click here to name the PRO

PROs include HRQoL/functional status, symptom/symptom burden, experience with care, health-related behaviors. (A PRO-based performance measure is not a survey instrument. Data may be collected using a survey instrument to construct a PRO measure.)

☐ Intermediate clinical outcome (*e.g., lab value*): Click here to name the intermediate outcome

☒ Process: Breast Cancer Screening

☐ Appropriate use measure: Click here to name what is being measured

☐ Structure: Click here to name the structure

☐ Composite: Click here to name what is being measured

1a.2 LOGIC MODEL Diagram or briefly describe the steps between the healthcare structures and processes (e.g., interventions, or services) and the patient's health outcome(s). The relationships in the diagram should be easily understood by general, non-technical audiences. Indicate the structure, process or outcome being measured.

2018 Submission

Females age 50-74 years >> screening for breast cancer is performed >> abnormal screening result >> evaluation and follow-up >> early detection and treatment >> improved length and/or quality of life

2014 Submission

Females age 50-74 years >> screening for breast cancer is performed >> results are evaluated >> results are positive for breast cancer >> treatment given >> improved length and/or quality of life

1a.3 Value and Meaningfulness: IF this measure is derived from patient report, provide evidence that the target population values the measured ***outcome, process, or structure*** and finds it meaningful. (Describe how and from whom their input was obtained.)

****RESPOND TO ONLY ONE SECTION BELOW -EITHER 1a.2, 1a.3 or 1a.4) ****

1a.2 FOR OUTCOME MEASURES including PATIENT REPORTED OUTCOMES - Provide empirical data demonstrating the relationship between the outcome (or PRO) to at least one healthcare structure, process, intervention, or service.

1a.3. SYSTEMATIC REVIEW(SR) OF THE EVIDENCE (for INTERMEDIATE OUTCOME, PROCESS, OR STRUCTURE PERFORMANCE MEASURES, INCLUDING THOSE THAT ARE INSTRUMENT-BASED) If the evidence is not based on a systematic review go to section 1a.4) If you wish to include more than one systematic review, add additional tables.

What is the source of the systematic review of the body of evidence that supports the performance measure? A systematic review is a scientific investigation that focuses on a specific question and uses explicit, prespecified scientific methods to identify, select, assess, and summarize the findings of similar but separate studies. It may include a quantitative synthesis (meta-analysis), depending on the available data. (IOM)

☐Clinical Practice Guideline recommendation (with evidence review)

☒ US Preventive Services Task Force Recommendation

☐Other systematic review and grading of the body of evidence (*e.g., Cochrane Collaboration, AHRQ Evidence Practice Center*)

☐Other

USPSTF Recommendation: • **Title** • **Author** • **Date** • **Citation, including page number** • **URL**	2018 Submission U.S. Preventive Services Task Force (USPSTF). 2016. Screening for Breast Cancer: U.S. Preventive Services Task Force Recommendation Statement. *Annals of Internal Medicine* 164(4) 279-296. doi: 10.7326/M15-2886. 2014 Submission U.S. Preventive Services Task Force. Screening for breast cancer: U.S. Preventive Services Task Force recommendation statement. Ann Intern Med 2009 Nov 17; 151(10):716-26, W-236. **URL:** http://www.uspreventiveservicestaskforce.org/uspstf/uspsbrca.htm
Quote the guideline or recommendation verbatim about the process, structure or intermediate outcome being measured. If not a guideline, summarize the conclusions from the SR.	2018 Submission The USPSTF recommends biennial screening mammography for women aged 50 to 74 years. Grade: B Recommendation The decision to start screening mammography in women prior to age 50 years should be an individual one. Women who place a higher value on the potential benefit than the potential harms may choose to begin biennial screening between the ages of 40 and 49 years. Grade: C Recommendation The USPSTF concludes that the current evidence is insufficient to assess the balance of benefits and harms of screening mammography in women aged 75 years or older. Grade: I Recommendation

	The USPSTF concludes that the current evidence is insufficient to assess the benefits and harms of digital breast tomosynthesis (DBT) as a primary screening method for breast cancer. Grade: I Recommendation The USPSTF concludes that the current evidence is insufficient to assess the balance of benefits and harms of adjunctive screening for breast cancer using breast ultrasonography, magnetic resonance imaging, DBT, or other methods in women identified to have dense breasts on an otherwise negative screening mammogram. Grade: I Recommendation 2014 Submission The USPSTF recommends biennial screening mammography for women aged 50 to 74 years. (B Recommendation) The decision to start regular, biennial screening mammography before the age of 50 years should be an individual one and take patient context into account, including the patient's values regarding specific benefits and harms. (C Recommendation) The USPSTF concludes that the current evidence is insufficient to assess the additional benefits and harms of screening mammography in women 75 years or older. (I Statement) The USPSTF recommends against teaching breast self-examination. (D Recommendation) The USPSTF concludes that the current evidence is insufficient to assess the additional benefits and harms of clinical breast examination beyond screening mammography in women 40 years or older. (I Statement) The USPSTF concludes that the current evidence is insufficient to assess the additional benefits and harms of either digital mammography or magnetic resonance imaging (MRI) instead of film mammography as screening modalities for breast cancer. (I Statement)
Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	2018 Submission The USPSTF concludes with moderate certainty that the net benefit of screening mammography in women aged 50 to 74 years is moderate. For a general population of women aged 40 to 49 years, there is moderate certainty that the net benefit of screening mammography in the general population of women. For women age 75 years and older, there is insufficient evidence on mammography screening and the balance of benefits and harms cannot be determined. The USPSTF concludes that the evidence on DBT as a primary screening modality for breast cancer is insufficient, and the balance of benefits and harms cannot be determined. The USPSTF concludes that the evidence on adjunctive screening for breast cancer using breast ultrasound, MRI, DBT, or other methods in women identified to have dense breasts on an otherwise negative

	screening mammogram is insufficient, and the balance of benefits and harms cannot be determined. 2014 Submission In the analytic framework, the USPSTF addressed in Key Question 1a whether screening with mammography (film or digital) or MRI decrease breast cancer mortality among women age 40-49 years and 70 and older. For this question, the USPSTF used seven studies in their meta-analysis, all rated *Fair*. The USPSTF grades the quality of the overall evidence for a service on a 3-point scale (good, fair, poor). A *Fair* rating means evidence is sufficient to determine effects on health outcomes, but the strength of the evidence is limited by the number, quality, or consistency of the individual studies, generalizability to routine practice, or indirect nature of the evidence on health outcomes.
Provide all other grades and definitions from the evidence grading system	2018 Submission N/A 2014 Submission *Good*: Evidence includes consistent results from well-designed, well-conducted studies in representative populations that directly assess effects on health outcomes. *Poor*: Evidence is insufficient to assess the effects on health outcomes because of limited number or power of studies, important flaws in their design or conduct, gaps in the chain of evidence, or lack of information on important health outcomes.
Grade assigned to the **recommendation** with definition of the grade	2018 Submission Grade B: The USPSTF recommends the service. There is high certainty that the net benefit is moderate or there is moderate certainty that the net benefit is moderate to substantial. Grade C: The USPSTF recommends selectively offering or providing this service to individual patients based on professional judgment and patient preferences. There is at least moderate certainty that the net benefit is small. Grade I: The USPSTF concludes that the current evidence is insufficient to assess the balance of benefits and harms of the service. Evidence is lacking, of poor quality, or conflicting, and the balance of benefits and harms cannot be determined. 2014 Submission The measure is based on a guideline to screen women age 50-74 years biennially, which is a grade B recommendation (Grade B: The USPSTF recommends the service. There is high certainty that the net benefit is moderate or there is moderate certainty that the net benefit is moderate to substantial).
Provide all other grades and definitions from the recommendation grading system	2018 Submission Grade A: The USPSTF recommends the service. There is high certainty that the net benefit is substantial.

	Grade D: The USPSTF recommends against the service. There is moderate or high certainty that the service has no net benefit or that the harms outweigh the benefits. 2014 Submission A Recommendation: The USPSTF recommends the service. There is high certainty that the net benefit is substantial. C Recommendation: Clinicians may provide this service to selected patients depending on individual circumstances. However, for most individuals without signs or symptoms there is likely to be only a small benefit from this service. D Recommendation: The USPSTF recommends against the service. There is moderate or high certainty that the service has no net benefit or that the harms outweigh the benefits. I Statement: The USPSTF concludes that the current evidence is insufficient to assess the balance of benefits and harms of the service. Evidence is lacking, of poor quality, or conflicting, and the balance of benefits and harms cannot be determined.
Body of evidence: • Quantity – how many studies? • Quality – what type of studies?	2018 Submission The evidence report supporting this guideline outlines the quantity and quality of evidence (Nelson et al 2016). Key Question 1: What is the effectiveness of routine mammography screening in reducing breast cancer–specific and all–cause mortality, and how does it differ by age, risk factor, and screening interval? • 65 studies (including 8 RCTs) of fair quality assessed breast cancer mortality in relationship to screening Key Question 2: What is the effectiveness of routine mammography screening in reducing the incidence of advanced breast cancer and treatment-related morbidity, and how does it differ by age, risk factors, and screening interval? • 20 studies (including 8 RCTs) of fair quality assessed incidence of advanced breast cancer in relationship to screening Key Question 3: How does the effectiveness of routine breast cancer screening in reducing breast cancer–specific and all-cause mortality vary by different screening modality? • No studies of tomosynthesis, ultrasound, or MRI addressed this question. Key Question 4: How does the effectiveness of routine breast cancer screening in reducing the incidence of advanced breast cancer and treatment-related morbidity vary by different screening modality? • 2 case studies of unknown quality compared digital mammography versus tomosynthesis and digital mammography reported detection rates by cancer stage using various categories of cancer staging.

	Key Question 5: What are the harms of routine mammography screening, and how do they differ by age, risk factor, and screening interval? • 53 studies (including meta-analyses, reviews, modeling studies, observational studies and a surveillance analysis of good and fair quality) assessed overdiagnosis, impact of false-positive and false-negative screening results on women, radiation exposure and incidence of pain, discomfort or distress after screening. Key Question 6: How do the harms of routine breast cancer screening vary by different screening modality? • 6 observational studies of fair or unknown quality compared false-positive recall rates of screening for breast cancer using mammography and tomosynthesis, or clinical breast exam compared with mammography alone. Nelson HD, Cantor A, Humphrey L, Fu R, Pappas M, Daeges M, Griffin J. Screening for Breast Cancer: A Systematic Review to Update the 2009 U.S. Preventive Services Task Force Recommendation. Evidence Synthesis No. 124. AHRQ Publication No. 14-05201-EF-1. Rockville, MD: Agency for Healthcare Research and Quality; 2016. 2014 Submission The 2009 evidence review included a meta-analysis of 7 randomized controlled trials of mammography screening. A modeling study estimated the benefits/harms of screening of different screening scenarios. The USPSTF Quality Rating for studies used in the meta analysis was fair.
Estimates of benefit and consistency across studies	2018 Submission The USPSTF found adequate evidence that mammography screening reduces breast cancer mortality in women aged 40 to 74 years. The number of breast cancer deaths averted increases with age; women aged 40 to 49 years benefit the least and women aged 60 to 69 years benefit the most. Age is the most important risk factor for breast cancer, and the increased benefit observed with age is at least partly due to the increase in risk. Women aged 40 to 49 years who have a first-degree relative with breast cancer have a risk for breast cancer similar to that of women aged 50 to 59 years without a family history. Direct evidence about the benefits of screening mammography in women aged 75 years or older is lacking. The USPSTF found inadequate evidence on the benefits and harms of DBT as a primary screening method for breast cancer. Similarly, the USPSTF found inadequate evidence on the benefits and harms of adjunctive screening for breast cancer using breast ultrasonography, MRI, DBT, or other methods in women identified to have dense breasts on an otherwise negative screening mammogram. In both cases, while there is some information about the accuracy of these methods, there is no information on the effects of their use on health outcomes, such as breast cancer incidence, mortality, or overdiagnosis rates.

	2014 Submission There is convincing evidence that screening with film mammography reduces breast cancer mortality, with a greater absolute reduction for women aged 50 to 74 years than for women aged 40 to 49 years. The strongest evidence for the greatest benefit is among women aged 60 to 69 years. Among women 75 years or older, evidence of benefits of mammography is lacking.
What harms were identified?	2018 Submission The USPSTF found adequate evidence that screening for breast cancer with mammography results in harms for women aged 40 to 74 years. The most important harm is the diagnosis and treatment of noninvasive and invasive breast cancer that would otherwise not have become a threat to a woman's health, or even apparent, during her lifetime (that is, overdiagnosis and overtreatment). False-positive results are common and lead to unnecessary and sometimes invasive follow-up testing, with the potential for psychological harms (such as anxiety). False-negative results (that is, missed cancer) also occur and may provide false reassurance. Radiation-induced breast cancer and resulting death can also occur, although the number of both of these events is predicted to be low. The USPSTF found inadequate evidence on the benefits and harms of DBT as a primary screening method for breast cancer. Similarly, the USPSTF found inadequate evidence on the benefits and harms of adjunctive screening for breast cancer using breast ultrasonography, MRI, DBT, or other methods in women identified to have dense breasts on an otherwise negative screening mammogram. In both cases, while there is some information about the accuracy of these methods, there is no information on the effects of their use on health outcomes, such as breast cancer incidence, mortality, or overdiagnosis rates. 2014 Submission Harms associated with screening for breast cancer include unnecessary imaging tests and biopsies in women without cancer and psychological harms and inconvenience due to false-positive screening results. Additional harms include treatment of cancer that would not become clinically apparent during a woman's lifetime (overdiagnosis) and the harms of unnecessary earlier treatment of breast cancer that would have become clinically apparent but would not have shortened a woman's life. Radiation exposure (from radiologic tests), although a minor concern, is also a consideration. The USPSTF determined that adequate evidence suggests that the overall harms associated with mammography are moderate for every age group considered. However, false-positive results are more common for women aged 40 to 49 years, whereas overdiagnosis is a greater concern for women in the older age groups. Results from randomized controlled trials show that screening mammography can help reduce the number of deaths from breast cancer, especially for those over age 50. The USPSTF noted with moderate certainty that the net benefits of screening mammography in women aged 50 to 74 years were at least moderate, and that the greatest benefits were seen in women aged 60 to 69 years. Thus, the harms did not outweigh the benefits in this age group.

Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	2018 Submission We are not aware of any published studies since the systematic review that would impact the recommendation. There are other clinical guidelines that recommend the use of digital breast tomosynthesis as a primary screening method for breast cancer, which we summarize in the section below. 2014 Submission N/A

1a.4 OTHER SOURCE OF EVIDENCE

If source of evidence is NOT from a clinical practice guideline, USPSTF, or systematic review, please describe the evidence on which you are basing the performance measure.

2018 Submission

The National Comprehensive Cancer Network (NCCN) publishes statements of evidence and expert consensus of currently accepted approaches to treatment. The American College of Radiology (ACR) publishes Appropriateness Criteria® for radiology procedures. NCCN's 2017 breast cancer screening clinical practice guideline and ACR's 2017 appropriateness criteria for breast cancer screening recommend the use of digital breast tomosynthesis for primary screening of breast cancer.

2014 Submission

N/A

1a.4.1 Briefly SYNTHESIZE the evidence that supports the measure. A list of references without a summary is not acceptable.

2018 Submission

NCCN states: "Combined use of digital mammography (two-dimensional, 2D) in conjunction with digital breast tomosynthesis (DBT) appears to improve cancer detection and reduce false-positive call-back rates. Tomosynthesis allows acquisition of three-dimensional (3D) data using a moving x-ray and digital detector. These data are reconstructed using computer algorithms to generate thin sections of images. The combined use of 2D and DBT results in double the radiation exposure compared with mammography alone. However, this increase in radiation dose falls below limits of radiation set by the U.S. Food and Drug Administration for standard mammography. The radiation dose can be minimized by newer tomosynthesis techniques that create a synthetic 2D image, which may obviate the need for a conventional digital image."

ACR states: "Digital breast tomosynthesis (DBT) can address some of the limitations encountered with standard mammographic views. In addition to planar images, DBT allows for creation and viewing of thin-section reconstructed images that may decrease the lesion-masking effect of overlapping normal tissue and reveal the true nature of potential false-positive findings without the need for recall. Several studies confirm that in a screening setting, the cancer detection rate is increased with use of DBT compared with 2-D mammography

alone. Additionally, the rate of recall for benign findings (false-positives) can be decreased. Some authors found these advantages to be especially pronounced in women under age 50, in those with dense breasts, and with lesion types including spiculated masses and asymmetries. Interpretation time for DBT images is greater than for standard mammography. Additionally, dose is increased if standard 2-D images are obtained in addition to DBT images. However, synthesized reconstructed images (a virtual planar image created from the tomographic dataset) may replace the need for a 2-D correlative view; current data suggest that these synthetic images perform as well as standard full-field digital images. DBT is almost always performed as part of an examination that also includes digital mammography. The digital mammography part of the examination may be in the form of traditional projection mammography or synthesized image from the DBT data."

2014 Submission

N/A

1a.4.2 What process was used to identify the evidence?

2018 Submission

NCCN: The development of the NCCN Guidelines is an ongoing and iterative process, which is based on a critical review of the best available evidence and derivation of recommendations by a multidisciplinary panel of experts in the field of cancer. Prior to the annual update of the Guidelines, an electronic search of the PubMed database, provided by the U.S. National Library of Medicine, is performed to obtain key literature published since the previous Guidelines update. The PubMed database was chosen as it remains the most widely used resource for medical literature and indexes only peer-reviewed biomedical literature. Articles from additional sources (e.g., e-publications ahead of print, meeting abstracts) deemed as relevant to the Guidelines may be included in the literature review process.

ACR: Appropriateness criteria are based on expert consensus and evidence review. A literature search was conducted in December 2015 and updated on March 2016 to identify additional evidence published since the ACR Appropriateness Criteria® Breast Cancer Screening topic was finalized. 379 articles were found. Twenty-four articles were added to the bibliography. The remaining articles were not used due to either poor study design, the articles were not relevant or generalizable to the topic, the results were unclear, misinterpreted, or biased, or the articles were already cited in the original bibliography. The author added 27 citations from bibliographies, websites, or books that were not found in the new literature search.

2014 Submission

N/A

1a.4.3. Provide the citation(s) for the evidence.

2018 Submission

National Comprehensive Cancer Network (NCCN). 2017. "Breast Cancer Screening and Diagnosis." (April 13, 2018). Guideline available at: https://www.nccn.org/professionals/physician_gls/pdf/breast-screening.pdf

American College of Radiology (ACR). 2017. "*ACR Appropriateness Criteria®: Breast Cancer Screening*." (April 13, 2018). Guideline available at: https://acsearch.acr.org/docs/70910/Narrative/

2014 Submission

N/A

NATIONAL QUALITY FORUM

Measure Information

This document contains the information submitted by measure developers/stewards, but is organized according to NQF's measure evaluation criteria and process. The item numbers refer to those in the submission form but may be in a slightly different order here. In general, the item numbers also reference the related criteria (e.g., item 1b.1 relates to sub criterion 1b).

Brief Measure Information

NQF #: 2372
Corresponding Measures:
De.2. **Measure Title:** Breast Cancer Screening
Co.1.1. **Measure Steward:** National Committee for Quality Assurance
De.3. **Brief Description of Measure:** Percentage of women 50-74 years of age who had a mammogram to screen for breast cancer
1b.1. **Developer Rationale:** This measure assesses screening for breast cancer using mammography, which can prevent or detect early breast cancer, as well as reduce deaths from breast cancer. Early detection of breast cancer by mammography may also allow for a greater range of treatment options, including less-aggressive surgery and less-invasive therapy.

S.4. **Numerator Statement:** Women who received a mammogram to screen for breast cancer.
S.6. **Denominator Statement:** Women 50-74 years of age.
S.8. **Denominator Exclusions:** This measure excludes women with a history of bilateral mastectomy. The measure also excludes patients who use hospice services or are enrolled in an institutional special needs plan or living long-term in an institution any time during the measurement year.

De.1. **Measure Type:** Process
S.17. **Data Source:** Claims, Electronic Health Data
S.20. **Level of Analysis:** Health Plan, Integrated Delivery System

IF Endorsement Maintenance – Original Endorsement Date: Sep 18, 2014 **Most Recent Endorsement Date:** Sep 18, 2014

IF this measure is included in a composite, NQF Composite#/title:

IF this measure is paired/grouped, NQF#/title:

De.4. **IF PAIRED/GROUPED, what is the reason this measure must be reported with other measures to appropriately interpret results?** N/A

1. Evidence, Performance Gap, Priority – Importance to Measure and Report

Extent to which the specific measure focus is evidence-based, important to making significant gains in healthcare quality, and improving health outcomes for a specific high-priority (high-impact) aspect of healthcare where there is variation in or overall less-than-optimal performance. ***Measures must be judged to meet all sub criteria to pass this criterion and be evaluated against the remaining criteria.***

1a. Evidence to Support the Measure Focus – See attached Evidence Submission Form
2._Evidence_Form_USPSTF_BCS.docx
1a.1 For Maintenance of Endorsement: Is there new evidence about the measure since the last update/submission?
Do not remove any existing information. If there have been any changes to evidence, the Committee will consider the new evidence. Please use the most current version of the evidence attachment (v7.1). Please use red font to indicate updated evidence.
Yes

1b. Performance Gap
Demonstration of quality problems and opportunity for improvement, i.e., data demonstrating:

- considerable variation, or overall less-than-optimal performance, in the quality of care across providers; and/or
- Disparities in care across population groups.

1b.1. Briefly explain the rationale for this measure *(e.g., how the measure will improve the quality of care, the benefits or improvements in quality envisioned by use of this measure)*
If a COMPOSITE (e.g., combination of component measure scores, all-or-none, any-or-none), SKIP this question and answer the composite questions.
This measure assesses screening for breast cancer using mammography, which can prevent or detect early breast cancer, as well as reduce deaths from breast cancer. Early detection of breast cancer by mammography may also allow for a greater range of treatment options, including less-aggressive surgery and less-invasive therapy.

1b.2. Provide performance scores on the measure as specified (current and over time) at the specified level of analysis. *(This is required for maintenance of endorsement. Include mean, std dev, min, max, interquartile range, scores by decile. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities include.) This information also will be used to address the sub-criterion on improvement (4b1) under Usability and Use.*
The following health-plan level data are collected through HEDIS and reflect the most recent years of measurement for this measure. Performance data are summarized at the health plan level and summarized by mean, standard deviation, minimum health plan performance, maximum health plan performance and performance at 10th, 25th, 50th, 75th, and 90th percentile. Data are stratified by year and product line (i.e. commercial, Medicaid, Medicare).

Commercial HMO/PPO Rate
YEAR | N | MEAN | STDEV| MIN |10TH | 25TH | 50TH| 75TH| 90TH | MAX
2015 | 410 | 71.9% | 5.9% | 52.0% | 64.8% | 68.1% | 71.1% | 75.8% | 80.5% | 89.6%
2016 | 426 | 71.4% | 6.5% | 37.2% | 64.5% | 67.7% | 71.0% | 75.4% | 80.2% | 88.2%
2017 | 420 | 71.4% | 7.0% | 13.2% | 64.0% | 68.0% | 71.2% | 75.7% | 79.8% | 88.0%

Medicaid HMO Rate
YEAR | N | MEAN | STDEV| MIN | 10TH | 25TH | 50TH| 75TH| 90TH | MAX
2015 | 170 | 58.77%| 09.75% | 36.70% | 45.99% | 51.59% | 58.37% | 66.02% | 71.32% | 87.88%
2016 | 201 | 58.51%| 10.09% | 17.78% | 47.38% | 52.28% | 58.10% | 65.06% | 71.44% | 88.51%
2017 | 257 | 58.87%| 09.24% | 30.56% | 48.00% | 52.71% | 59.02% | 65.51% | 70.29% | 87.92%

Medicare HMO/PPO Rate
YEAR | N | MEAN | STDEV| MIN | 10TH | 25TH | 50TH| 75TH| 90TH | MAX
2015 | 425 | 71.0% | 11.3% | 09.8% | 60.5% | 66.8% | 71.7% | 78.4% | 82.9% | 92.1%
2016 | 405 | 72.4% | 09.8% | 14.3% | 62.0% | 67.4% | 72.7% | 79.2% | 83.3% | 91.9%
2017 | 431 | 72.2% | 09.6% | 18.4% | 61.4% | 67.0% | 73.0% | 78.8% | 82.9% | 91.1%

In 2016, HEDIS measures covered 114.2 million commercial health plan beneficiaries, 47 million Medicaid beneficiaries and 17.6 million Medicare beneficiaries. Below is a description of the denominator for this measure. It includes the number of health plans included in HEDIS data collection and the mean eligible population for the measure across health plans.

Breast Cancer Screening – commercial
YEAR | N Plans | Median Denominator Size per plan
2015 | 410 | 7,740
2016 | 426 | 7,510
2017 | 420 | 7,552

Breast Cancer Screening – Medicaid
YEAR | N Plans | Median Denominator Size per plan
2015 | 172 | 1,642
2016 | 202 | 1,419
2017 | 258 | 2,439

Breast Cancer Screening – Medicare
YEAR | N Plans | Median Denominator Size per plan
2015 | 425 | 2,173

2016 | 405 | 2,297
2017 | 431 | 1,890

1b.3. If no or limited performance data on the measure as specified is reported in 1b2, then provide a summary of data from the literature that indicates opportunity for improvement or overall less than optimal performance on the specific focus of measurement.
N/A

1b.4. Provide disparities data from the measure as specified (current and over time) by population group, e.g., by race/ethnicity, gender, age, insurance status, socioeconomic status, and/or disability. *(This is required for maintenance of endorsement. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included.) For measures that show high levels of performance, i.e., "topped out", disparities data may demonstrate an opportunity for improvement/gap in care for certain sub-populations. This information also will be used to address the sub-criterion on improvement (4b1) under Usability and Use.*
HEDIS data are stratified by type of insurance (e.g. Commercial, Medicaid, Medicare). While not specified in the measure, health plans can stratify by demographic variables, such as race/ethnicity or socioeconomic status, in order to assess the presence of health care disparities. The HEDIS Health Plan Measure Set contains two measures that can assist with stratification to assess health care disparities. The Race/Ethnicity Diversity of Membership and the Language Diversity of Membership measures were designed to promote standardized methods for collecting these data and follow Office of Management and Budget and Institute of Medicine guidelines for collecting and categorizing race/ethnicity and language data. In addition, the NCQA Multicultural Health Care Distinction Program outlines standards for collecting, storing and using race/ethnicity and language data to assess health care disparities. Starting in 2019, Medicare Advantage plans will report this measure stratified by low-income subsidy/dual eligibility and disability status, which are proxies for low socioeconomic status.

1b.5. If no or limited data on disparities from the measure as specified is reported in 1b.4, then provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement. Include citations. Not necessary if performance data provided in 1b.4
Studies have identified disparities in breast cancer screening based on race, ethnicity, education and income. One study found that mammography use in 2006 was 65 percent among white women and 59 percent among black women (Njai et al 2011). Additionally, African American women are more likely than white women to have longer intervals between screening mammograms, which may lead to an increase in later-stage cancer diagnoses (CDC 2012). Between 2010 and 2014, breast cancer mortality for African American women was 41 percent higher than white women (Richardson et al 2016); one potential contributing factor to this health disparity is access to mammography screening services (Rust et al 2015).National survey data also show that women who have attained lower degrees of education, lack health insurance coverage or have lower socioeconomic status are less likely to have recently had a mammogram (National Center for Health Statistics 2015).

National Center for Health Statistics. 2015."Health, United States, 2015: With Special Feature on Racial and Ethnic Health Disparities". http://www.cdc.gov/nchs/data/hus/hus15.pdf#070 (Accessed Mar 13, 2018).

Njai, R., Siegel, P. Z., Miller, J. W., & Liao, Y. 2011. "Misclassification of Survey Responses and Black-White Disparity in Mammography Use, Behavioral Risk Factor Surveillance System, 1995-2006." Preventing Chronic Disease 8(3), A59.

Richardson, L.C., Henley, S.J., Miller, J.W., Massetti, G., and Thomas, C.C. 2016. "Patterns and Trends in Age-Specific Black-White Differences in Breast Cancer Incidence and Mortality – United States, 1999–2014." Morbidity and Mortality Weekly Report (MMWR) 65(40); 1093-1098. (November 30, 2016) http://www.cdc.gov/mmwr/volumes/65/wr/mm6540a1.htm.

Rust, G., Zhang, S., Malhotra, K., Reese, L., McRoy, L., Baltrus, P., Caplan, L. and Levine, R. 2015. "Paths to Health Equity – Local Area Variation in Progress Toward Eliminating Breast Cancer Mortality Disparities, 1990–2009." Cancer 121(16): 2765-2774. (December 1, 2016) doi: 10.1002/cncr.29405.

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.
- 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] 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] 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]

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]

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; [14,15] 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] 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*

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.

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

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

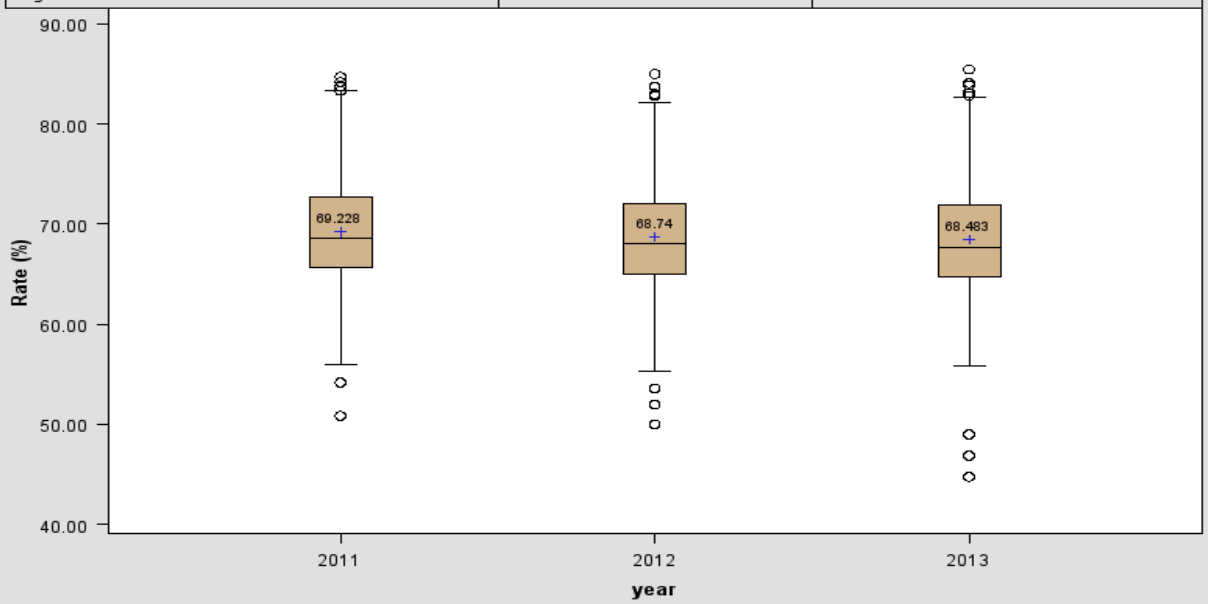
HEDIS 2013 Variation in Performance across Health Plans
Note: results are from the measure specified for women 40-69 years, the most recent data available. The measure was updated in 2013 primarily to assess women 50-74 years.

	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.

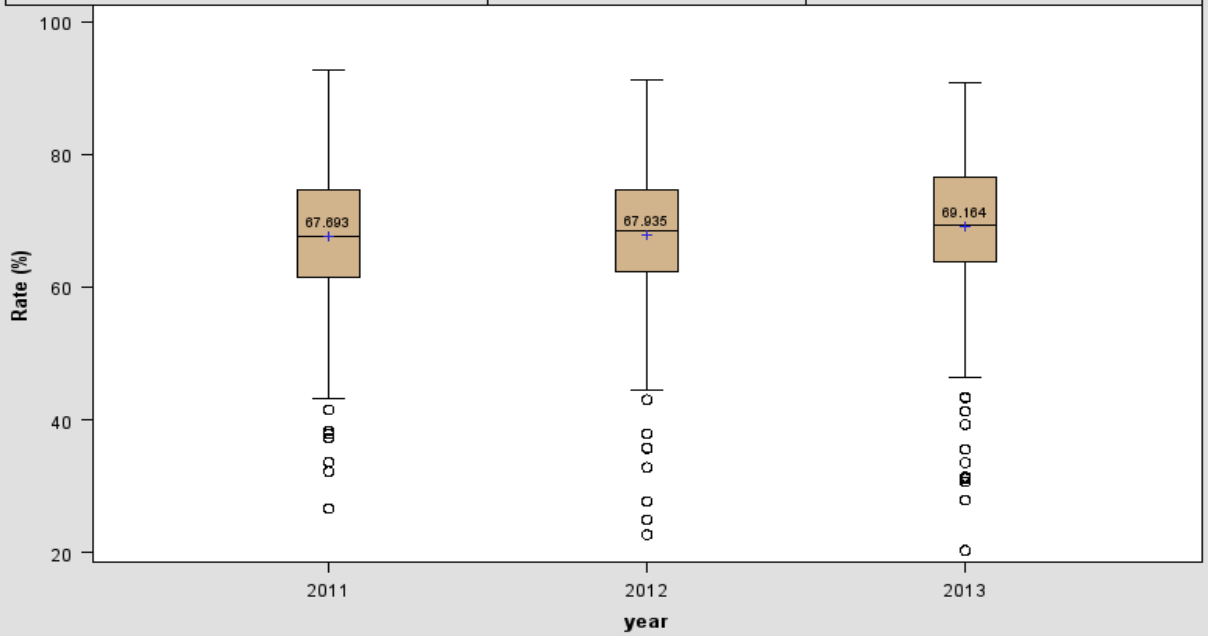
Commercial Graph for Breast Cancer Screening from 2011-2013

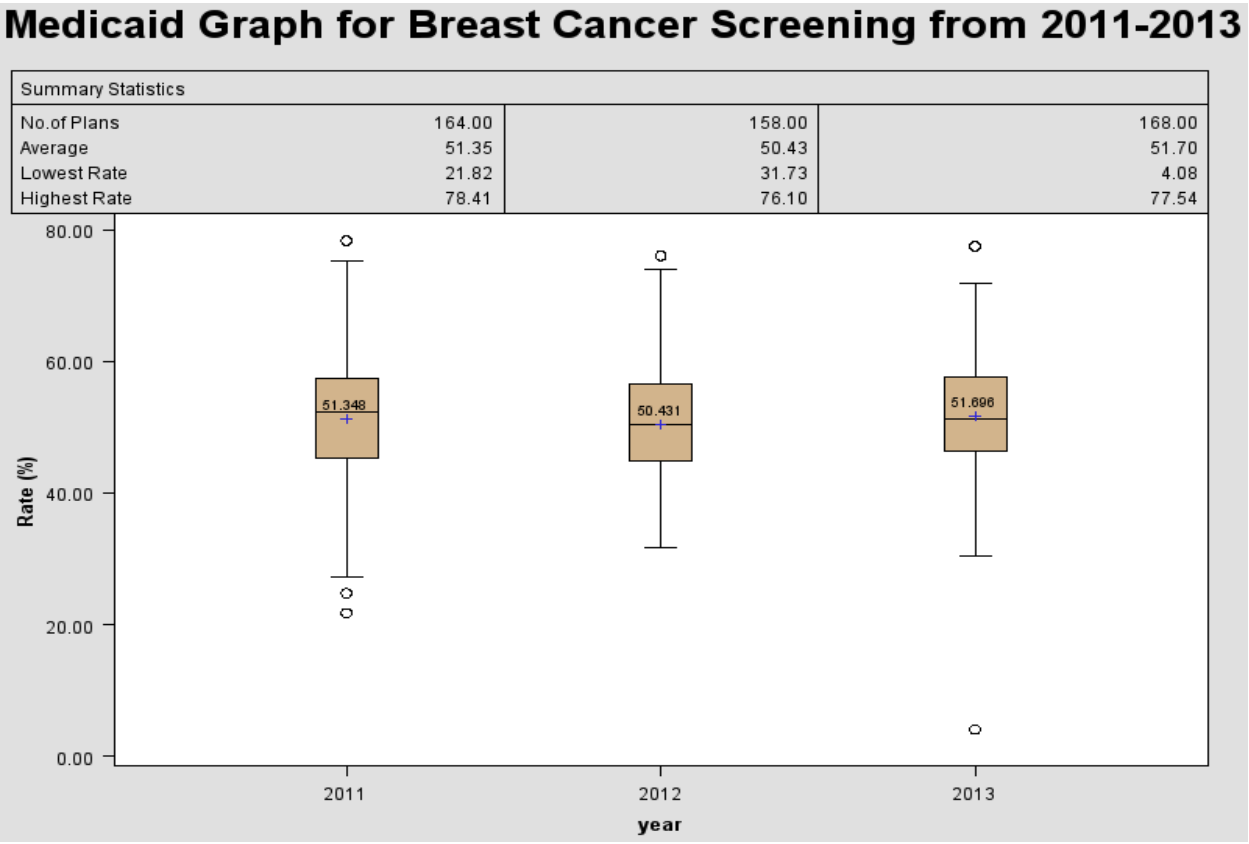
Summary Statistics			
No.of Plans	412.00	408.00	419.00
Average	69.23	68.74	68.48
Lowest Rate	50.87	50.05	44.78
Highest Rate	84.73	85.01	85.46



Medicare Graph for Breast Cancer Screening from 2011-2013

Summary Statistics			
No.of Plans	424.00	452.00	478.00
Average	67.69	67.93	69.16
Lowest Rate	26.67	22.73	20.37
Highest Rate	92.65	91.16	90.71





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.

2. Reliability and Validity—Scientific Acceptability of Measure Properties

Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. ***Measures must be judged to meet the sub criteria for both reliability and validity to pass this criterion and be evaluated against the remaining criteria.***

2a.1. Specifications The measure is well defined and precisely specified so it can be implemented consistently within and across organizations and allows for comparability. eMeasures should be specified in the Health Quality Measures Format (HQMF) and the Quality Data Model (QDM).

De.5. **Subject/Topic Area** *(check all the areas that apply):*
Cancer, Cancer : Breast

De.6. **Non-Condition Specific***(check all the areas that apply):*
Primary Prevention, Screening

De.7. **Target Population Category** *(Check all the populations for which the measure is specified and tested if any):*

S.1. **Measure-specific Web Page** *(Provide a URL link to a web page specific for this measure that contains current detailed specifications including code lists, risk model details, and supplemental materials. Do not enter a URL linking to a home page or to general information.)*
N/A

S.2a. **If this is an eMeasure**, HQMF specifications must be attached. Attach the zipped output from the eMeasure authoring tool (MAT) - if the MAT was not used, contact staff. (Use the specification fields in this online form for the plain-language description of the specifications)
This is not an eMeasure **Attachment:**

S.2b. **Data Dictionary, Code Table, or Value Sets** *(and risk model codes and coefficients when applicable) must be attached. (Excel or csv file in the suggested format preferred - if not, contact staff)*
Attachment **Attachment:** 2372_Breast_Cancer_Screening_Value_Sets-636594894640541618.xlsx

S.2c. Is this an instrument-based measure (i.e., data collected via instruments, surveys, tools, questionnaires, scales, etc.)? Attach copy of instrument if available.
No, this is not an instrument-based measure **Attachment:**

S.2d. Is this an instrument-based measure (i.e., data collected via instruments, surveys, tools, questionnaires, scales, etc.)? Attach copy of instrument if available.
Not an instrument-based measure

S.3.1. **For maintenance of endorsement:** Are there changes to the specifications since the last updates/submission. If yes, update the specifications for S1-2 and S4-22 and explain reasons for the changes in S3.2.
Yes

S.3.2. **For maintenance of endorsement,** please briefly describe any important changes to the measure specifications since last measure update and explain the reasons.
Since the last NQF review, digital breast tomosynthesis was added as an acceptable breast cancer screening method in order to account for the use of this method by women with clinical indications. This change was reviewed by stakeholder groups, vetted through a public comment period, and approved by our committees.

S.4. **Numerator Statement** *(Brief, narrative description of the measure focus or what is being measured about the target population, i.e., cases from the target population with the target process, condition, event, or outcome) DO NOT include the rationale for the measure.*
IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm (S.14).
Women who received a mammogram to screen for breast cancer.

S.5. **Numerator Details** *(All information required to identify and calculate the cases from the target population with the target process, condition, event, or outcome such as definitions, time period for data collection, specific data collection items/responses,*

code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)
IF an OUTCOME MEASURE, describe how the observed outcome is identified/counted. Calculation of the risk-adjusted outcome should be described in the calculation algorithm (S.14).
One or more mammograms any time on or between October 1 two years prior to the measurement year and December 31 of the measurement year.

Notes:
(1) This measure assesses the use of imaging to detect early breast cancer in women. Because the measure denominator does not remove women at higher risk of breast cancer, all types and methods of mammograms (screening, diagnostic, film, digital or digital breast tomosynthesis) qualify for numerator compliance. MRIs, ultrasounds or biopsies do not count toward the numerator; although they may be indicated for evaluating women at higher risk for breast cancer or for diagnostic purposes, they are performed as an adjunct to mammography and do not themselves count toward the numerator.
(2) The numerator time frame is 27 months. NCQA allows for a 3-month leeway, a method used for other HEDIS measures (as determined on a per-measure basis), in recognition of the logistics of referrals and scheduling and to avoid potential overuse of screening. This time frame was recommended by our expert advisory panels and approved by our Committee on Performance Measurement, which oversees measures used in the HEDIS Health Plan Measures Set.

See attached code value sets.

S.6. Denominator Statement *(Brief, narrative description of the target population being measured)*
Women 50-74 years of age.

S.7. Denominator Details *(All information required to identify and calculate the target population/denominator such as definitions, time period for data collection, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b.)*
IF an OUTCOME MEASURE, describe how the target population is identified. Calculation of the risk-adjusted outcome should be described in the calculation algorithm (S.14).
Women 52-74 years as of the end of the measurement year (December 31).
Note: this denominator statement captures women age 50-74 years; it is structured to account for the look-back period for mammograms.

S.8. Denominator Exclusions *(Brief narrative description of exclusions from the target population)*
This measure excludes women with a history of bilateral mastectomy. The measure also excludes patients who use hospice services or are enrolled in an institutional special needs plan or living long-term in an institution any time during the measurement year.

S.9. Denominator Exclusion Details *(All information required to identify and calculate exclusions from the denominator such as definitions, time period for data collection, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b.)*
Exclude patients with bilateral mastectomy any time during the member's history through December 31 of the measurement year. Any of the following meet criteria for bilateral mastectomy:
1) Bilateral mastectomy (Bilateral Mastectomy Value Set)
2) Unilateral mastectomy (Unilateral Mastectomy Value Set) with a bilateral modifier (Bilateral Modifier Value Set)
3) Two unilateral mastectomies (Unilateral Mastectomy Value Set) with service dates 14 days or more apart
4) History of bilateral mastectomy (History of Bilateral Mastectomy Value Set)
5) Any combination of codes that indicate a mastectomy on both the left and right side on the same or different dates of service. Left mastectomy includes any of the following: unilateral mastectomy (Unilateral Mastectomy Value Set) with a left-side modifier (Left Modifier Value Set) same claim; or absence of the left breast (Absence of Left Breast Value Set); or left unilateral mastectomy (Unilateral Mastectomy Left Value Set). Right Mastectomy includes any of the following: unilateral mastectomy (Unilateral Mastectomy Value Set) with a right-side modifier (Right Modifier Value Set) same claim; or absence of the right breast (Absence of Right Breast Value Set); or right unilateral mastectomy (Unilateral Mastectomy Right Value Set).

Exclude patients who use hospice services any time during the measurement year (Hospice Value Set).

Exclude patients 65 and older who are enrolled in an institutional SNP or living long-term in an institution at any time during the measurement year.

S.10. **Stratification Information** *(Provide all information required to stratify the measure results, if necessary, including the stratification variables, definitions, specific data collection items/responses, code/value sets, and the risk-model covariates and coefficients for the clinically-adjusted version of the measure when appropriate – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format with at S.2b.)*
N/A

S.11. **Risk Adjustment Type** (Select type. Provide specifications for risk stratification in measure testing attachment)
No risk adjustment or risk stratification
If other:

S.12. **Type of score:**
Rate/proportion
If other:

S.13. **Interpretation of Score** *(Classifies interpretation of score according to whether better quality is associated with a higher score, a lower score, a score falling within a defined interval, or a passing score)*
Better quality = Higher score

S.14. **Calculation Algorithm/Measure Logic** (*Diagram or describe the calculation of the measure score as an ordered sequence of steps including identifying the target population; exclusions; cases meeting the target process, condition, event, or outcome; time period for data, aggregating data; risk adjustment; etc.)*
Step 1. Determine the eligible population: identify women 52-74 years of age by the end of the measurement year.
Step 2. Search for an exclusion: history of bilateral mastectomy; or use of hospice services during the measurement year; or patients 65 and older who are enrolled in an institutional SNP or living long-term in an institution any time during measurement year. Exclude these patients from the eligible population.
Step 3. Determine numerator: the number of patients who received one or more mammograms any time on or between October 1 two years prior to the measurement year and December 31 of the measurement year.
Step 4. Calculate the rate.

S.15. **Sampling** *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*
IF an instrument-based performance measure (e.g., PRO-PM), identify whether (and how) proxy responses are allowed.
N/A

S.16. **Survey/Patient-reported data** *(If measure is based on a survey or instrument, provide instructions for data collection and guidance on minimum response rate.)*
Specify calculation of response rates to be reported with performance measure results.
N/A

S.17. **Data Source** *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*
If other, please describe in S.18.
Claims, Electronic Health Data

S.18. **Data Source or Collection Instrument** *(Identify the specific data source/data collection instrument (e.g. name of database, clinical registry, collection instrument, etc., and describe how data are collected.)*
IF instrument-based, identify the specific instrument(s) and standard methods, modes, and languages of administration.
This measure is based on administrative claims collected in the course of providing care to health plan members. NCQA collects the Healthcare Effectiveness Data and Information Set (HEDIS) data for this measure directly from Health Management Organizations and Preferred Provider Organizations via NCQA's online data submission system.

S.19. **Data Source or Collection Instrument** *(available at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
No data collection instrument provided

S.20. **Level of Analysis** *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*
Health Plan, Integrated Delivery System

S.21. **Care Setting** *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*
Outpatient Services
If other:

S.22. **COMPOSITE Performance Measure** - Additional Specifications *(Use this section as needed for aggregation and weighting rules, or calculation of individual performance measures if not individually endorsed.)*
N/A

2. Validity – See attached Measure Testing Submission Form
3._Testing_Form_BCS.docx

2.1 For maintenance of endorsement
Reliability testing: If testing of reliability of the measure score was not presented in prior submission(s), has reliability testing of the measure score been conducted? If yes, please provide results in the Testing attachment. Please use the most current version of the testing attachment (v7.1). Include information on all testing conducted (prior testing as well as any new testing); use red font to indicate updated testing.
Yes

2.2 For maintenance of endorsement
Has additional empirical validity testing of the measure score been conducted? If yes, please provide results in the Testing attachment. Please use the most current version of the testing attachment (v7.1). Include information on all testing conducted (prior testing as well as any new testing); use red font to indicate updated testing.
Yes

2.3 For maintenance of endorsement
Risk adjustment: For outcome, resource use, cost, and some process measures, risk-adjustment that includes social risk factors is not prohibited at present. Please update sections 1.8, 2a2, 2b1,2b4.3 and 2b5 in the Testing attachment and S.140 and S.11 in the online submission form. NOTE: These sections must be updated even if social risk factors are not included in the risk-adjustment strategy. You MUST use the most current version of the Testing Attachment (v7.1) -- older versions of the form will not have all required questions.
No - This measure is not risk-adjusted

3. Feasibility

Extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

3a. Byproduct of Care Processes

For clinical measures, the required data elements are routinely generated and used during care delivery (e.g., blood pressure, lab test, diagnosis, medication order).

3a.1. **Data Elements Generated as Byproduct of Care Processes.**
Generated or collected by and used by healthcare personnel during the provision of care (e.g., blood pressure, lab value, diagnosis, depression score), Coded by someone other than person obtaining original information (e.g., DRG, ICD-9 codes on claims)
If other:

3b. Electronic Sources

The required data elements are available in electronic health records or other electronic sources. If the required data are not in electronic health records or existing electronic sources, a credible, near-term path to electronic collection is specified.

3b.1. **To what extent are the specified data elements available electronically in defined fields** (*i.e., data elements that are needed to compute the performance measure score are in defined, computer-readable fields*) Update this field for **maintenance of endorsement**.
ALL data elements are in defined fields in electronic claims

3b.2. **If ALL the data elements needed to compute the performance measure score are not from electronic sources, specify a credible, near-term path to electronic capture, OR provide a rationale for using other than electronic sources.** For **maintenance of endorsement**, if this measure is not an eMeasure (eCQM), please describe any efforts to develop an eMeasure (eCQM).
N/A

3b.3. **If this is an eMeasure, provide a summary of the feasibility assessment in an attached file or make available at a measure-specific URL. Please also complete and attach the NQF Feasibility Score Card.**
Attachment:

3c. Data Collection Strategy

Demonstration that the data collection strategy (e.g., source, timing, frequency, sampling, patient confidentiality, costs associated with fees/licensing of proprietary measures) can be implemented (e.g., already in operational use, or testing demonstrates that it is ready to put into operational use). For eMeasures, a feasibility assessment addresses the data elements and measure logic and demonstrates the eMeasure can be implemented or feasibility concerns can be adequately addressed.

3c.1. **Required for maintenance of endorsement. Describe difficulties (as a result of testing and/or operational use of the measure) regarding data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, time and cost of data collection, other feasibility/implementation issues.**
IF instrument-based, consider implications for both individuals providing data (patients, service recipients, respondents) and those whose performance is being measured.
This measure is specified for administrative data, which has been found to be highly feasible. Further, NCQA conducts an independent audit of all HEDIS collection and reporting processes in order to verify that HEDIS specifications are met. NCQA has developed a precise, standardized methodology for verifying the integrity of HEDIS collection and calculation processes through a two-part program consisting of an overall information systems capabilities assessment followed by an evaluation of the MCO's ability to comply with HEDIS specifications. NCQA-certified auditors using standard audit methodologies will help enable purchasers to make more reliable "apples-to-apples" comparisons between health plans. The HEDIS Compliance Audit addresses the following functions:
1) information practices and control procedures
2) sampling methods and procedures
3) data integrity
4) compliance with HEDIS specifications
5) analytic file production
6) reporting and documentation

In addition to the HEDIS Audit, NCQA provides a system to allow "real-time" feedback from measure users. Through our Policy Clarification Support System, NCQA responds immediately to technical questions regarding measures in order to promote consistent implementation of the measure.

Input from NCQA auditing and the Policy Clarification Support System informs the annual updating of all HEDIS measures including updating value sets and clarifying the specifications. Measures are re-evaluated on a periodic basis and when there is a significant change in evidence. During re-evaluation, information from NCQA auditing and Policy Clarification Support System (as well as through stakeholder advisory panels) informs evaluation of the scientific soundness and feasibility of the measure.

3c.2. **Describe any fees, licensing, or other requirements to use any aspect of the measure as specified** *(e.g., value/code set, risk model, programming code, algorithm)*.
Broad public use and dissemination of these measures is encouraged and NCQA has agreed with NQF that noncommercial uses do not require the consent of the measure developer. Use by health care physicians in connection with their own practices is not commercial use. Commercial use of a measure requires the prior written consent of NCQA. As used herein, "commercial use" refers to any sale, license or distribution of a measure for commercial gain, or incorporation of a measure into any product or service that is sold, licensed or distributed for commercial gain, even if there is no actual charge for inclusion of the measure.

4. Usability and Use

Extent to which potential audiences (e.g., consumers, purchasers, providers, policy makers) are using or could use performance results for both accountability and performance improvement to achieve the goal of high-quality, efficient healthcare for individuals or populations.

4a. Accountability and Transparency

Performance results are used in at least one accountability application within three years after initial endorsement and are publicly reported within six years after initial endorsement (or the data on performance results are available). If not in use at the time of initial endorsement, then a credible plan for implementation within the specified timeframes is provided.

4.1. Current and Planned Use

NQF-endorsed measures are expected to be used in at least one accountability application within 3 years and publicly reported within 6 years of initial endorsement in addition to performance improvement.

Specific Plan for Use	Current Use (for current use provide URL)
	Public Reporting NCQA Health Plan Ratings https://reportcards.ncqa.org/#/health-plans/list NCQA Quality Compass http://www.ncqa.org/tabid/177/Default.aspx CMS Qualified Health Plan (QHP) Quality Rating System (QRS) https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/QualityInitiativesGenInfo/Downloads/2018_QRS_and_QHP_Enrollee_Survey_Technical_Guidance_20171004_508.pdf California's Value Based Pay for Performance Program http://www.iha.org/our-work/accountability/value-based-p4p CMS Quality Payment Program https://qpp.cms.gov/ NCQA State of Health Care Quality http://www.ncqa.org/report-cards/health-plans/state-of-health-care-quality NCQA Health Plan Ratings https://reportcards.ncqa.org/#/health-plans/list NCQA Quality Compass http://www.ncqa.org/tabid/177/Default.aspx CMS Qualified Health Plan (QHP) Quality Rating System (QRS) https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/QualityInitiativesGenInfo/Downloads/2018_QRS_and_QHP_Enrollee_Survey_Technical_Guidance_20171004_508.pdf California's Value Based Pay for Performance Program http://www.iha.org/our-work/accountability/value-based-p4p CMS Quality Payment Program https://qpp.cms.gov/ NCQA State of Health Care Quality http://www.ncqa.org/report-cards/health-plans/state-of-health-care-quality Payment Program CMS Medicare Star Rating Program https://www.medicare.gov/find-a-plan/questions/home.aspx CMS Medicaid Adult Core Set https://www.medicaid.gov/medicaid/quality-of-care/performance-measurement/adult-core-set/index.html California's Value Based Pay for Performance Program http://www.iha.org/our-work/accountability/value-based-p4p CMS Quality Payment Program https://qpp.cms.gov/ CMS Medicare Star Rating Program https://www.medicare.gov/find-a-plan/questions/home.aspx CMS Medicaid Adult Core Set

	https://www.medicaid.gov/medicaid/quality-of-care/performance-measurement/adult-core-set/index.html California's Value Based Pay for Performance Program http://www.iha.org/our-work/accountability/value-based-p4p CMS Quality Payment Program https://qpp.cms.gov/ Regulatory and Accreditation Programs NCQA Health Plan Accreditation http://www.ncqa.org/tabid/123/Default.aspx NCQA Health Plan Accreditation http://www.ncqa.org/tabid/123/Default.aspx Quality Improvement (external benchmarking to organizations) NCQA Quality Compass http://www.ncqa.org/tabid/177/Default.aspx NCQA State of Health Care Quality http://www.ncqa.org/report-cards/health-plans/state-of-health-care-quality

4a1.1 **For each CURRENT use, checked above (update for <u>maintenance of endorsement</u>), provide:**

- Name of program and sponsor
- Purpose
- Geographic area and number and percentage of accountable entities and patients included
- Level of measurement and setting

CALIFORNIA VALUE BASED PAY FOR PERFORMANCE PROGRAM: This measure is used in the California P4P program, which is the largest non-governmental physician incentive program in the United States. Founded in 2001, it is managed by the Integrated Healthcare Association (IHA) on behalf of ten health plans representing 9 million insured persons. IHA reports results on approximately 35,000 physicians in 200 physician organizations.

CMS MEDICARE ADVANTAGE STAR RATING PROGRAM: This measure is included in the composite Medicare Advantage Star Rating. CMS calculates a Star Rating (1-5) for all Medicare Advantage health plans based on 53 performance measures. Medicare beneficiaries can view the star rating and individual measure scores on the CMS Plan Compare website. The Star Rating is also used to calculate bonus payments to health plans with excellent performance. The Medicare Advantage Plan Rating program covers 11.5 million Medicare beneficiaries in 455 health plans across all 50 states.

CMS MEDICAID ADULT CORE SET: There are a core set of health quality measures for Medicaid-enrolled adults. The Medicaid Adult Core Set was identified by the Centers of Medicare & Medicaid (CMS) in partnership with the Agency for Healthcare Research and Quality (AHRQ). The data collected from these measures will help CMS to better understand the quality of health care that adults enrolled in Medicaid receive nationally. Beginning in January 2014 and every three years thereafter, the Secretary is required to report to Congress on the quality of care received by adults enrolled in Medicaid. Additionally, beginning in September 2014, state data on the adult quality measures will become part of the Secretary's annual report on the quality of care for adults enrolled in Medicaid.

CMS QUALIFIED HEALTH PLAN (QHP) QUALITY RATING SYSTEM (QRS): This measure is used in the Qualified Health Plan (QHP) Quality Rating System, which provides comparable information to consumers about the quality of health care services and QHP enrollee experience offered in the Marketplaces.

CMS QUALITY PAYMENT PROGRAM: This measure is used in the Quality Payment Program (QPP) which is a reporting program that uses a combination of incentive payments and payment adjustments to promote reporting of quality information by eligible professionals (EPs).

NCQA STATE OF HEALTH CARE QUALITY REPORT: This measure is publicly reported nationally and by geographic regions in the NCQA State of Health Care annual report. This annual report published by NCQA summarizes findings on quality of care. In 2012, the report included measures on 11.5 Medicare Advantage beneficiaries in 455 Medicare Advantage health plans, 99.4 million members in 404 commercial health plans, and 14.3 million Medicaid beneficiaries in 136 plans across 50 states.

NCQA HEALTH PLAN RATINGS/REPORT CARDS: This measure is used to calculate health plan ratings, which are reported in Consumer Reports and on the NCQA website. These rankings are based on performance on HEDIS measures among other factors. In 2012, a total of 455 Medicare Advantage health plans, 404 commercial health plans, and 136 Medicaid health plans across 50 states were included in the rankings.

NCQA HEALTH PLAN ACCREDITATION: This measure is used in scoring for accreditation of Medicare Advantage Health Plans. In 2012, a total of 170 Medicare Advantage health plans were accredited using this measure among others covering 7.1 million Medicare beneficiaries and 336 commercial health plans covering 87 million lives. Health plans are scored based on performance compared to benchmarks.

NCQA QUALITY COMPASS: This measure is used in Quality Compass which is an indispensable tool used for selecting health plans, conducting competitor analysis, examining quality improvement and benchmarking plan performance. Provided in this tool is the ability to generate custom reports by selecting plans, measures, and benchmarks (averages and percentiles) for up to three trended years. Results in table and graph formats offer simple comparison of plans' performance against competitors or benchmarks.

4a1.2. **If not currently publicly reported OR used in at least one other accountability application (e.g., payment program, certification, licensing) what are the reasons?** (*e.g., Do policies or actions of the developer/steward or accountable entities restrict access to performance results or impede implementation?*)
N/A

4a1.3. **If not currently publicly reported OR used in at least one other accountability application, provide a credible plan for implementation within the expected timeframes -- any accountability application within 3 years and publicly reported within 6 years of initial endorsement.** (*Credible plan includes the specific program, purpose, intended audience, and timeline for implementing the measure within the specified timeframes. A plan for accountability applications addresses mechanisms for data aggregation and reporting.*)
N/A

4a2.1.1. **Describe how performance results, data, and assistance with interpretation have been provided to those being measured or other users during development or implementation.**
How many and which types of measured entities and/or others were included? If only a sample of measured entities were included, describe the full population and how the sample was selected.
During development, NCQA receives input from those reporting and using measures through several multistakeholder advisory panels and a broad public comment posting. For this particular measure, the clinical advisory panel included several representatives from health plans and users such as federal policymakers and consumers. We also sought input from our standing Technical Measurement Advisory Panel, which includes representatives from health plans and other users and advises NCQA on feasibility and other potential implementation issues. During implementation, health plans that report HEDIS calculate their rates and know their performance when submitting to NCQA. NCQA publicly reports rates across all plans and also creates benchmarks in order to help plans understand how they perform relative to other plans. Public reporting and benchmarking are effective quality improvement methods.

4a2.1.2. **Describe the process(es) involved, including when/how often results were provided, what data were provided, what educational/explanatory efforts were made, etc.**
NCQA publishes HEDIS results annually in our Quality Compass tool. NCQA also presents data at various conferences and webinars. For example, at the annual HEDIS Update and Best Practices Conference, NCQA presents results from all new measures' first year of implementation or analyses from measures that have changed significantly. NCQA also regularly provides technical assistance on measures through its Policy Clarification Support System, as described in Section 3c.1.

4a2.2.1. **Summarize the feedback on measure performance and implementation from the measured entities and others described in 4d.1.**
Describe how feedback was obtained.
NCQA measures are evaluated regularly using a consensus-based process to consider input from multiple stakeholders, including but not limited to entities being measured. We use several methods to obtain input, including vetting of the measure with several multi-stakeholder advisory panels, public comment posting, and review of questions submitted to the Policy Clarification Support System. This information enables NCQA to comprehensively assess a measure's adherence to the HEDIS Desirable Attributes of Relevance, Scientific Soundness and Feasibility.

4a2.2.2. **Summarize the feedback obtained from those being measured.**
During development and reevaluation, those reporting and using measures reported the measure continues to relevant and important for quality improvement and accountability. Questions received through the Policy Clarification Support system have generally sought clarification about the mammography screening methods that satisfy the measure numerator. During a recent public comment session, a majority of comments from measured entities supported updates to the measure to align with the latest clinical recommendations.

4a2.2.3. **Summarize the feedback obtained from other users**
This measure has been deemed a priority measure by NCQA and other entities, as illustrated by its use in programs such as the Medicare Advantage Star Rating program and the Medicaid Adult Core Set.

4a2.3. **Describe how the feedback described in 4a2.2.1 has been considered when developing or revising the measure specifications or implementation, including whether the measure was modified and why or why not.**
During the measure's last major update, feedback obtained through the mechanisms described in 4a2.2.1 informed how we revised the measure to include digital breast tomosynthesis as a new screening method.

Improvement
Progress toward achieving the goal of high-quality, efficient healthcare for individuals or populations is demonstrated. If not in use for performance improvement at the time of initial endorsement, then a credible rationale describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.

4b1. **Refer to data provided in 1b but do not repeat here. Discuss any progress on improvement (trends in performance results, number and percentage of people receiving high-quality healthcare; Geographic area and number and percentage of accountable entities and patients included.)**
If no improvement was demonstrated, what are the reasons? If not in use for performance improvement at the time of initial endorsement, provide a credible rationale that describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.
From 2011 to 2017, average performance rates increased by about six percentage points for Medicare and Medicaid plans and were steady for commercial plans. Over the past three years, average performance rates slightly increased for Medicare plans and were stable for Commercial and Medicaid plans. In 2017, average performance was about 71% for Commercial and Medicare plans, and 59% for Medicaid plans.

There continues to be variation between the 10th and 90th percentile, suggesting room for improvement. In 2017, commercial plans in the 10th percentile had a rate of 64% compared to 80% for plans in the 90th percentile; Medicare plans in the 10th percentile had a rate of 61% compared to 83% for plans in the 90th percentile; and Medicaid plans in the 10th percentile had a rate of 48% compared to 70% for plans in the 90th percentile.

4b2. Unintended Consequences
The benefits of the performance measure in facilitating progress toward achieving high-quality, efficient healthcare for individuals or populations outweigh evidence of unintended negative consequences to individuals or populations (if such evidence exists).

4b2.1. **Please explain any unexpected findings (positive or negative) during implementation of this measure including unintended impacts on patients.**
One potential unintended consequence of the Breast Cancer Screening measure is too-frequent screening. The U.S. Preventive Services Task Force recommends biennial screening in women age 50-74. Feedback from our advisory panel indicated that, in an effort to meet the two-year requirement, women often are encouraged to seek screening earlier than the two-year mark. In order to address potential over-screening, NCQA adjusted the numerator time frame to 27 months, providing a three-month leeway to account for the logistics of scheduling and receiving a mammogram.

4b2.2. **Please explain any unexpected benefits from implementation of this measure.**
Benefits from implementation of the measure were as expected.

5. Comparison to Related or Competing Measures

If a measure meets the above criteria and there are endorsed or new related measures (either the same measure focus or the same target population) or competing measures (both the same measure focus and the same target population), the measures are compared to address harmonization and/or selection of the best measure.

5. Relation to Other NQF-endorsed Measures
Are there related measures (conceptually, either same measure focus or target population) or competing measures (conceptually both the same measure focus and same target population)? If yes, list the NQF # and title of all related and/or competing measures.
Yes

5.1a. **List of related or competing measures (selected from NQF-endorsed measures)**
0508 : Diagnostic Imaging: Inappropriate Use of "Probably Benign" Assessment Category in Screening Mammograms
0509 : Diagnostic Imaging: Reminder System for Screening Mammograms

5.1b. **If related or competing measures are not NQF endorsed please indicate measure title and steward.**
N/A

5a. Harmonization of Related Measures
The measure specifications are harmonized with related measures;
OR
The differences in specifications are justified

5a.1. **If this measure conceptually addresses EITHER the same measure focus OR the same target population as NQF-endorsed measure(s):**
Are the measure specifications harmonized to the extent possible?
Yes

5a.2. **If the measure specifications are not completely harmonized, identify the differences, rationale, and impact on interpretability and data collection burden.**
Both related measures have a different focus than our health plan screening measure. NQF #0509 Reminder System for Mammograms is intended to encourage implementation of reminder systems for future mammograms. NQF #0508 Inappropriate Use of "Probably Benign" Assessment Category focuses on accurate documentation of mammogram results. Both measures are also specified at the clinician level rather than the health plan level.

5b. Competing Measures
The measure is superior to competing measures (e.g., is a more valid or efficient way to measure);
OR
Multiple measures are justified.

5b.1. **If this measure conceptually addresses both the same measure focus and the same target population as NQF-endorsed measure(s):**
Describe why this measure is superior to competing measures (e.g., a more valid or efficient way to measure quality); OR provide a rationale for the additive value of endorsing an additional measure. (Provide analyses when possible.)
N/A

Appendix

A.1 Supplemental materials may be provided in an appendix. All supplemental materials (such as data collection instrument or methodology reports) should be organized in one file with a table of contents or bookmarks. If material pertains to a specific submission form number, that should be indicated. Requested information should be provided in the submission form and required attachments. There is no guarantee that supplemental materials will be reviewed.
No appendix **Attachment:**

Contact Information

Co.1 **Measure Steward (Intellectual Property Owner):** National Committee for Quality Assurance
Co.2 **Point of Contact:** Bob, Rehm, rehm@ncqa.org, 202-955-1728-
Co.3 **Measure Developer if different from Measure Steward:** National Committee for Quality Assurance

Co.4 **Point of Contact:** Bob, Rehm, rehm@ncqa.org, 202-955-1728-

Additional Information

Ad.1 **Workgroup/Expert Panel involved in measure development**

Provide a list of sponsoring organizations and workgroup/panel members' names and organizations. Describe the members' role in measure development.

The NCQA Breast Cancer Screening Measurement Advisory Panels advised NCQA during measure development. They evaluated the way staff specified the measure, reviewed field test results, and assessed NCQA's overall desirable attributes of Relevance, Scientific Soundness, and Feasibility. The advisory panel consisted of a balanced group of experts. In addition to this advisory panel, we vetted the measure with a host of other stakeholders, as is our process. Thus, our measures are the result of consensus from a broad and diverse group of stakeholders.

2014 BREAST CANCER SCREENING MEASUREMENT ADVISORY PANEL MEMBERS
Kathy Coltin, Harvard Pilgrim Health Care
Laura Esserman, University of California, San Francisco
Lisa Latts, formally at WellPoint, Inc.
Nancy Lee, U.S. Department of Health & Human Services
Dorothy Mann, UW School of Medicine and School of Public Health and Community Medicine
Melissa McNeil, University of Pittsburgh
Ellen Stovall, National Coalition for Cancer Survivorship
Richard Wender, Thomas Jefferson University

2017 BREAST CANCER SCREENING MEASUREMENT ADVISORY PANEL MEMBERS
Joanne Armstrong, Aetna
Laura Esserman, University of California, San Francisco
Sandra Finestone, Association of Cancer Patient Educators
David Larsen, Intermountain Healthcare
Melissa McNeil, University of Pittsburgh, UPMC
Robert Smith, American Cancer Society

2017 GERIATRIC MEASUREMENT ADVISORY PANEL MEMBERS
Arlene Bierman, AHRQ
Patricia Bomba, Excellus BlueCross BlueShield
Jennie Chin Hansen, American Geriatrics Society (Retired)
Joyce Dubow, Public Member
Peter Hollmann, Brown University
Steven Phillips, Geriatric Specialty Care
Wade Aubry, UCSF Institute for Health Policy
Jane Sung, AARP
Eric Tangalos, Mayo Clinic
Dirk Wales, Cigna HealthSpring
Neil Wenger, UCLA
Nicole Brandt, UMD Pharmacy
Karen Nichols, Amerihealth Caritas
Gustavo Ferrer, Aventura Hospital
Jeff Kelman, CMS
Joan Weiss, HHS

2017 COMMITTEE ON PERFORMANCE MEASUREMENT
Bruce Bagley, Independent Consultant
Andrew Baskin, Aetna
Jonathan D. Darer, Medicalis
Helen Darling, Strategic Advisor on Health Benefits & Health Care
Kate Goodrich, Centers for Medicare and Medicaid Services
David Grossman, Kaiser Permanente
Christine S. Hunter, US Office of Personnel Management
Jeffrey Kelman, Centers for Medicare & Medicaid Services

Nancy Lane, Vanderbilt University Medical Center
Bernadette Loftus, The Permanente Medical Group
Adrienne Mims, Alliant Quality
Amanda Parsons, Montefiore Health System
Eric C. Schneider, The Commonwealth Fund
Marcus Thygeson, Blue Shield of California
JoAnn Volk, Georgetown University Center on Health Insurance Reforms

2017 TECHNICAL MEASUREMENT ADVISORY PANEL MEMBERS
Andy Amster, Kaiser Permanente
Jennifer Brudnicki, Geisinger Health Plan
Lindsay Cogan, New York State Department of Health
Kathy Coltin, Independent Consultant
Mike Farina, MVP Healthcare
Marissa Finn, CIGNA HealthCare
Scott Fox, Independence Blue Cross
Carlos Hernandez, CenCal Health
Harmon Jordan, Westat
Virginia Raney, Center for Medicaid and CHIP Services
Lynne Rothney-Kozlak, Rothney-Kozlak Consulting, LLC
Laurie Spoll, Aetna

Measure Developer/Steward Updates and Ongoing Maintenance
Ad.2 **Year the measure was first released:** 1995
Ad.3 **Month and Year of most recent revision:** 04, 2018
Ad.4 **What is your frequency for review/update of this measure?** Approximately every 3 years, sooner if the clinical guidelines have changed significantly.
Ad.5 **When is the next scheduled review/update for this measure?** 12, 2019

Ad.7 **Disclaimers:** These performance Measures are not clinical guidelines and do not establish a standard of medical care, and have not been tested for all potential applications.

THE MEASURES AND SPECIFICATIONS ARE PROVIDED "AS IS" WITHOUT WARRANTY OF ANY KIND.

Ad.8 **Additional Information/Comments:** NCQA Notice of Use. Broad public use and dissemination of these measures is encouraged and NCQA has agreed with NQF that noncommercial uses do not require the consent of the measure developer. Use by health care physicians in connection with their own practices is not commercial use. Commercial use of a measure requires the prior written consent of NCQA. As used herein, "commercial use" refers to any sale, license, or distribution of a measure for commercial gain, or incorporation of a measure into any product or service that is sold, licensed, or distributed for commercial gain, even if there is no actual charge for inclusion of the measure.