NATIONAL QUALITY FORUM

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 #: 0509

Corresponding Measures:

De.2. Measure Title: Diagnostic Imaging: Reminder System for Screening Mammograms

Co.1.1. Measure Steward: American College of Radiology

De.3. Brief Description of Measure: Percentage of patients undergoing a screening mammogram whose information is entered into a reminder system with a target due date for the next mammogram

1b.1. Developer Rationale: Although screening mammograms can reduce breast cancer mortality by 20-35% in women aged 40 years and older, recent evidence shows that only 72% of women are receiving mammograms based on current guideline recommendations. The use of patient reminders is associated with an increase in screening mammography. Encouraging the implementation of a reminder system could lead to an increase in mammography screening at appropriate intervals.

Any facility that uses less than annual frequency of screening, greatly increases the importance of attendance at each scheduled screening. Even with annual screening recommendations screening does not always occur biennially. This demonstrates the importance of systematic reminders and active patient outreach.

The purpose of screening is to minimize interval or false negative cancers, as these are failures of the screening process. The 2011 article by Bennett, Sellars and Moss (Ref 1)and an earlier work by Woodman, Threlfall and Boggis (ref 2) examine the effect of interval cancer rates (false negative cancers) by time since screen out to three years in the United Kingdom´s triennial screening program. The Interval cancer rates (false negative cases) increase over time (ref 1,2) and begin to approach incidence rates by the third year (ref 2). Thus screening at greater than 2 year intervals will likely have poor overall outcomes in reducing breast cancer mortality. These papers may also underestimate the rate of interval cancers (ref 1) so the actual rates may be higher.

Efforts to ensure regular screening are therefore necessary to eliminate any screening interval beyond 2 years.

1. Bennett RL, Sellars SJ, Moss SM. Interval cancers in the NHS breast cancer screening programme in England, Wales and Northern Ireland. British Journal of Cancer. 2011. 104: 571-577.

2. Woodman CBJ, Threlfall AG, Boggis CR et al. Is the three year breast screening interval too long? Occurrence of interval cancers in NHS breast screening programme's north western region. BMJ. 1995. 310:224-6

S.4. Numerator Statement: Patients whose information is entered into a reminder system with a target due date for the next mammogram

S.6. Denominator Statement: All patients undergoing a screening mammogram

S.8. Denominator Exclusions: Documentation of medical reason(s) for not entering patient information into a reminder system [(eg, further screening mammograms are not indicated, such as patients with a limited life expectancy, other medical reason(s)]

De.1. Measure Type: Structure

S.17. Data Source: Claims, Registry Data

S.20. Level of Analysis: Clinician : Individual

IF Endorsement Maintenance – Original Endorsement Date: Oct 28, 2008 **Most Recent Endorsement Date:** Oct 26, 2016

Preliminary Analysis: Maintenance of Endorsement

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**

Summary of prior review in 2016

- The developer provided a recommendation from the Community Preventive Services Task Force that recommends the use of client reminders to increase screening for breast and cervical cancers on the basis of strong evidence of effectiveness. **Level of Evidence: Recommended.**
    - The Task Force describes "Recommended" as: The systematic review of available studies provides strong or sufficient evidence that the intervention is effective. The categories of "strong" and "sufficient" evidence reflect the Task Force's degree of confidence that an intervention has beneficial effects. They do not directly relate to the expected magnitude of benefits. The categorization is based on several factors, such as study design, number of studies, and consistency of the effect across studies.
- The developer provided a systematic review (not graded) and a summary of the QQC demonstrating the effectiveness of reminder systems in increasing breast cancer screening by mammography.

Summary of prior review in 2008

- The evidence for this measure was based on guidelines recommendations from the American College of Radiology (ACR) for the performance of screening and diagnostic mammography and the American College of Radiology (ACR) Breast Imaging Reporting and Data System Atlas (BI-RADS® Atlas). The strength of evidence was not ranked.

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**

The developer provided an additional 2018 study of a 4-year randomized trial comparing three outreach interventions to promote screening mammography that reinforced the previous evidence: "A simple reminder call can increase screening mammogram adherence even when baseline adherence is high."

Exception to evidence

N/A

Question for the Committee:

- *The updated evidence is directionally consistent with and strengthens the underlying evidence since the last NQF endorsement review. Does the Committee agree the evidence basis for the measure has not changed and there is no need for repeat discussion and vote on Evidence?*

Guidance from the Evidence Algorithm

Measure a health outcome (Box 1) No →Process/structure measure with systematic review and grading of Task Force recommendation (Box 3) Yes → Summary of the quantity, quality, and consistency (QQC) of the body of evidence (Box 4) Yes → Box 5a) → HIGH

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 developers provided scores on this measure for 2015-2018; 79,450 physicians with at least 10 patients had a non-zero reporting rate.
- 7.4% of physicians did not meet the measure.
- The mean performance rate for the four years (2015-2018) was 95.69%. The developer provided the rates by year:
  - 2015, 88.0%
  - 2016, 94.4%
  - 2017, 96.4%
  - 2018, 95.6%
- The performance rate quartiles for 2015-2018 for physicians with at least 10 patients and performance rate >0 were:
  - 25th percentile: 99.5%
  - 50th percentile: 100%
  - 75th percentile: 100%

2016 Committee Review Data

- During the previous review: 47,866 physicians with at least 10 patients had a non-zero reporting rate.
- Across these physicians, 15.0% of physicians did not meet the measure.
- For the 3-year period 2012-2014 the mean performance rate was 85.0%.
    - 2012, 79.4%
    - 2013, 86.0%
    - 2014, 87.6%
- The performance rate quartiles for 2012-2014 for physicians with at least 10 patients and performance rate >0 were:
    - 25th percentile: 91.15%
    - 50th percentile: 100%
    - 75th percentile: 100%

Disparities

- The developer did not provide updated disparities data from the measure as specified nor a review of the literature.
- Previously, the developer stated that based on 2010 data from the National Health Interview Survey (NHIS) Asian race, low education status, recent immigrant status, and no regular source of medical care or no medical insurance were factors found to reduce the likelihood for a woman to receive a mammogram, but the developer did not provide disparities data related to the focus of this measure, i.e., a reminder system.

Questions for the Committee:

- Is there a gap in care and/or disparities that warrant a national performance measure? The mean performance for 2014-2018 was 95.69%. Should this measure be considered for Inactive Endorsement with Reserve Status?
- Since no disparities information is provided, are you aware of evidence that disparities exist in this area of healthcare (reminder system, not mammography per se)?

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

RATIONALE:

- The mean performance for 2014-2018 was 95.69%. No disparities data are provided.

Committee Pre-evaluation Comments:
Criteria 1: Importance to Measure and Report (including 1a, 1b, 1c)

1a. Evidence
Comments:

- Structure
- High
- US Preventive Services Task Force 2016 report on mammography and 2019 recommendations indicate strong evidence supporting the recommendations. In addition an NAP report indicates that reminder systems are recommended to increase adherence to mammography recommendations (NAP 2005). See online https://www.nap.edu/read/11308/chapter/7#165

1b. Performance Gap/Disparities
Comments:

- Yes, performance gap data on the measure is provided. To the extent possible, the performance measure data should be stratified by race, ethnicity, and primary language to assess disparities and initiate subsequent quality improvement activities addressing identified disparities, consistent with recent national efforts to standardize the collection of race and ethnicity data. Having said that, the developer did not provide updated disparities data from the measure as specified nor a review of the literature. Further, the developer did not provide disparities data related to the focus of this measure, i.e., a reminder system. Lastly, mean performance appears to be "topped-off?"
- I didn't see data on population subgroups. There are known disparities in cancer screening in general and in cancer survival. Sending reminders should be standard across a system. From the evidence cited they state that reminders do correlate with women getting mammograms, but no data was provided for subgroups. Do reminders impact disparities or are there other barriers factors underlying the disparities in screening. I agree with summary provided by NQF staff that the performance gap is low.
- The gap reported is very small and the figure reported was 7% not meeting the measure and approx 95% meeting the measure. The room for improvement is small. The measure does not provide a lot of value for encouraging improvement. Disparity information was not provided, so there might be some residual disparities that could respond to improvement efforts; but we do not know, absent the data on disparities.

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

Questions for the Committee regarding reliability:

- Do you have any concerns that the measure can be consistently implemented (i.e., are measure specifications adequate)?
- Reliability statistic has increased from 0.88 to 0.98 since the last evaluation. Is this increase likely the result of performance on the measure being topped out and so should be considered for Reserve Status?

Questions for the Committee regarding validity:

- Do you have any concerns regarding the validity of the measure (e.g., exclusions, risk-adjustment approach, etc.) or based on the empirical score-level testing? (For the current submission, the developer performed construct validity. It was unable to correlate the measure to an outcome, but compared this measure to two other process measures (NQF 509 Reminder System for Screening Mammograms and ACRad5: Screening Mammography Abnormal Interpretation (Recall Rate), hypothesizing that a good performance on this measure likely indicates physicians who follow guidelines are likely in practices with good systems for tracking patients/remind patients and that physicians who do well on ACRad5 do not unnecessarily recall patients, respectively.)
- In its assessment of meaningful differences (threat to validity), the developer reported that for claims, registry, or QCDR data there is "minor statistically significant and clinically meaningful differences?" To what degree does the Committee feel this threatens the validity of the measure?

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

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

Scientific Acceptability: Preliminary Analysis Form

Measure Number: 0509

Measure Title: Diagnostic Imaging: Reminder System for Screening Mammograms

Type of measure:

☐ **Process** ☐ **Process: Appropriate Use** ☒ **Structure** ☐ **Efficiency** ☐ **Cost/Resource Use**

☐ **Outcome** ☐ **Outcome: PRO-PM** ☐ **Outcome: Intermediate Clinical Outcome** ☐ **Composite**

Data Source:

☒ **Claims** ☐ **Electronic Health Data** ☐ **Electronic Health Records** ☐ **Management Data**
☐ **Assessment Data** ☐ **Paper Medical Records** ☐ **Instrument-Based Data** ☒ **Registry Data**
☐ **Enrollment Data** ☐ **Other**

Level of Analysis:

☐ **Clinician: Group/Practice** ☒ **Clinician: Individual** ☐ **Facility** ☐ **Health Plan**
☐ **Population: Community, County or City** ☐ **Population: Regional and State**
☐ **Integrated Delivery System** ☐ **Other**

Measure is:

☐ **New** ☒ **Previously endorsed (**NOTE: Empirical validity testing is expected at time of maintenance review; if not possible, justification is required.)

RELIABILITY: SPECIFICATIONS

1. **Are submitted specifications precise, unambiguous, and complete so that they can be consistently implemented?** ☒ **Yes** ☐ **No**

   Submission document: "MIF_xxxx" document, items S.1-S.22

   ***NOTE**: NQF staff will conduct a separate, more technical, check of eCQM specifications, value sets, logic, and feasibility, so no need to consider these in your evaluation.*

2. **Briefly summarize any concerns about the measure specifications.**
   - No concerns

RELIABILITY: TESTING

Submission document: "MIF_xxxx" document for specifications, testing attachment questions 1.1-1.4 and section 2a2

3. **Reliability testing level** ☒ **Measure score** ☐ **Data element** ☐ **Neither**
4. **Reliability testing was conducted with the data source and level of analysis indicated for this measure** ☒ **Yes** ☐ **No**
5. If score-level and/or data element reliability testing was NOT conducted or if the methods used were NOT appropriate, was **empirical VALIDITY testing** of patient-level data conducted?

   ☐ **Yes** ☐ **No**

6. **Assess the method(s) used for reliability testing**

   Submission document: Testing attachment, section 2a2.2

   - The beta-binomial method to assess the signal-to-noise ratio was conducted (N=79.450 physicians) .
   - The developer noted that reliability was estimated at two points: 1) at a minimum number of reporting events per physician and 2) at the average number of quality reporting events per physician. The minimum threshold of events was set at 10.
   - CMS physician-level claims, registry, and QCDR data was extracted for the relevant physician-level information.

7. **Assess the results of reliability testing**

   Submission document: Testing attachment, section 2a2.3

   The developer provided the following results for its empiric reliability testing:

   - The mean reliability for the data abstracted during 2012-2014 was 0.88.
   - The mean reliability for the data abstracted from 2015-2018 was 0.98.
   - The developer stated that a statistic 0.80 is considered very good reliability.

8. 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.

   Submission document: Testing attachment, section 2a2.2

   ☒ **Yes**

   ☐ **No**

   ☐ **Not applicable** (score-level testing was not performed)

9. Was the method described and appropriate for assessing the reliability of ALL critical data elements?

   Submission document: Testing attachment, section 2a2.2

   ☐ **Yes**

   ☐ **No**

   ☒ **Not applicable** (data element testing was not performed)

10. **OVERALL RATING OF RELIABILITY** (taking into account precision of specifications 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** (NOTE: Should rate LOW if you believe specifications are NOT precise, unambiguous, and complete or if testing methods/results are not adequate)

    ☐ **Insufficient** (NOTE: Should rate INSUFFICIENT if you believe you do not have the information you need to make a rating decision)

11. **Briefly explain rationale for the rating of OVERALL RATING OF RELIABILITY and any concerns you may have with the approach to demonstrating reliability.**

    Score-level testing was conducted. The developer notes that the reliability statistic of 0.98 by convention indicates high reliability (1.0 is "perfect").

VALIDITY: ASSESSMENT OF THREATS TO VALIDITY

12. **Please describe any concerns you have with measure exclusions.**

    Submission document: Testing attachment, section 2b2.

    - No exclusions.

13. **Please describe any concerns you have regarding the ability to identify meaningful differences in performance.**

    Submission document: Testing attachment, section 2b4.

    - For registry data, the performance rates did not show significant variation, with an interquartile range of 0%. There is no statistically significant difference in performance between the top and bottom quartile ($P<0.0001$ at alpha = 0.05). The developer stated this variation shows that there is minor statistically significant and clinically meaningful differences in rates across providers submitting registry information.
    - For the QCDR data, the performance rates did not show significant variation, with an interquartile range of 0%. There is no statistically significant difference in performance between the top and bottom quartile ($P<0.0001$ at alpha = 0.05). The developer stated this variation shows that there is minor statistically significant and clinically meaningful differences in rates across providers submitting QCDR information.

14. **Please describe any concerns you have regarding comparability of results if multiple data sources or methods are specified.**

    Submission document: Testing attachment, section 2b5.

    - This measure has only one set of specifications.

15. **Please describe any concerns you have regarding missing data.**

    Submission document: Testing attachment, section 2b6.

    The developer noted the following:

- CMS Medicare and Medicaid administrative data are considered highly valid and reliable, since they determine eligibility for enrollment and payment of services. Registry data may have some non-responders, as they are not required to submit all data to CMS. However, the developer further noted that the volume of patients (68,844,412) used in the registry data set greatly minimizes the risk of bias. It stated that, each year, CMS raises the amount of data required for submission in the MIPS program, which should assist with minimizing bias even more in the future.

16. **Risk Adjustment**

16a. **Risk-adjustment method** ☒ **None** ☐ **Statistical model** ☐ **Stratification**

16b. **If not risk-adjusted, is this supported by either a conceptual rationale or empirical analyses?**

☐ Yes ☐ No ☒ Not applicable

16c. **Social risk adjustment:**

16c.1 Are social risk factors included in risk model? ☐ Yes ☐ No ☒ Not applicable

16c.2 Conceptual rationale for social risk factors included? ☐ Yes ☐ No

16c.3 Is there a conceptual relationship between potential social risk factor variables and the measure focus? ☐ Yes ☐ No

16d.**Risk adjustment summary:**

16d.1 All of the risk-adjustment variables present at the start of care? ☐ Yes ☐ No

16d.2 If factors not present at the start of care, do you agree with the rationale provided for inclusion? ☐ Yes ☐ No

16d.3 Is the risk adjustment approach appropriately developed and assessed? ☐ Yes ☐ No

16d.4 Do analyses indicate acceptable results (e.g., acceptable discrimination and calibration) ☐ Yes ☐ No

16d.5.Appropriate risk-adjustment strategy included in the measure? ☐ Yes ☐ No

16e**. Assess the risk-adjustment approach**

For cost/resource use measures ONLY:

17. **Are the specifications in alignment with the stated measure intent?**

☐ **Yes** ☐ **Somewhat** ☐ **No (If "Somewhat" or "No", please explain)**

18. **Describe any concerns of threats to validity related to attribution, the costing approach, carve outs, or truncation (approach to outliers):**

VALIDITY: TESTING

19. **Validity testing level:** ☒ **Measure score** ☐ **Data element** ☐ **Both**

20. **Method of establishing validity of the measure score:**

☒ **Face validity**

☒ **Empirical validity testing of the measure score**

☐ **N/A (score-level testing not conducted)**

21. **Assess the method(s) for establishing validity**

Submission document: Testing attachment, section 2b2.2

- For the current submission, the developer performed construct validity. It was unable to correlate the measure to an outcome, but compared this measure to two other process measures (NQF 509 Reminder System for Screening Mammograms and ACRad5: Screening Mammography Abnormal Interpretation (Recall Rate), hypothesizing that a good performance on this measure likely indicates physicians who

follow guidelines are likely in practices with good systems for tracking patients/remind patients and that physicians who do well on ACRad5 do not unnecessarily recall patients, respectively.

- For its face validity testing (previous submissions), the developer noted: An expert panel was used to assess face validity of the measure. Panel consisted of members from the American College of Radiology Commission on Breast Imaging and the National Mammography Database.

22. **Assess the results(s) for establishing validity**

Submission document: Testing attachment, section 2b2.3

For empiric validity testing:

- For the current submission, the developer performed construct validity and calculated Pearson's correlation coefficients. It was unable to correlate the measure to an outcome, so used two process measures, as noted at Item 21. The developer did not find a relationship between the performance rates of the measures; Pearson's correlation coefficients all hover around 0, which is no relationship; of note, both NQF 508 and NQF 509 are nearly "topped out." The developer maintains the measure has high face validity.

For face validity, the developer noted (previous 2016 submission):

- The expert panel was asked for its level of agreement on the following statements and whether the measure remained valid based on existing and new evidence.
- The results of the expert panel responses showed agreement (80% or higher) on the following statements:
  - the measure demonstrates a high impact on health care and an opportunity for improvement in quality over time.
  - physicians who perform well on this measure demonstrate a higher level of quality than physicians who do not perform well on the measure.
  - this measure would increase awareness of appropriate use.
  - believed it would promote higher quality management and treatment.
- The results of the expert panel responses also showed that 54.5% believed this measure was complementary to the recall rate metric in Hospital Compare

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

Submission document: Testing attachment, section 2b1.

☒ **Yes**

☐ **No**

☐ **Not applicable** (score-level testing was not performed)

24. **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.*

Submission document*: Testing attachment, section 2b1.*

☐ **Yes**

☐ **No**

☒ **Not applicable** (data element testing was not performed)

25. **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** (NOTE: Should rate LOW if you believe that there are threats to validity and/or relevant threats to validity were not assessed OR if testing methods/results are not adequate)

☒ **Insufficient** (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.)

26. **Briefly explain rationale for rating of OVERALL RATING OF VALIDITY and any concerns you may have with the developers' approach to demonstrating validity.**

- For the current submission, the developer performed construct validity. It was unable to correlate the measure to an outcome, so used two process measures, as noted at Item 22. The developer did not find a relationship between the performance rates of the measures; Pearson's correlation coefficients all hover around 0, which is no relationship; of note, both NQF 508 and NQF 509 are nearly "topped out." The developer maintains the measure has high face validity.
- The developer's assessment of meaningful differences in performance demonstrate there is "no statistically significant difference in measure rates between the top and bottom quartile" and "minor statistically significant and clinically meaningful differences in performance."

Box 1. Potential threats to validity assessed >> meaningful differences, assessed, but no statistical difference between top and bottom quartile = NO >> INSUFFICIENT.

Box 1. Potential threats to validity assessed >> although no statistically significant difference between top and bottom quartile, developer states "there is minor statistically significant and clinically meaningful differences" =YES >> Box 2. Empirical Testing >> Box 5 Score Level Testing >> Box 6 Appropriate Method >> Box 7c >> No relationship identified so INSUFFICIENT/LOW.

FOR COMPOSITE MEASURES ONLY: Empirical analyses to support composite construction

27. **What is the level of certainty or confidence that the empirical analysis demonstrates that the component measures add value to the composite and that the aggregation and weighting rules are consistent with the quality construct?**

☐ **High**

☐ **Moderate**

☐ **Low**

☐ **Insufficient**

28. **Briefly explain rationale for rating of EMPIRICAL ANALYSES TO SUPPORT COMPOSITE CONSTRUCTION**

ADDITIONAL RECOMMENDATIONS

29. **If you have listed any concerns in this form, do you believe these concerns warrant further discussion by the multi-stakeholder Standing Committee? If so, please list those concerns below.**

- The measure, as specified, does not identify statistically significant differences between the top and bottom quartile. The lack of meaningful differences is a threat to validity.

Committee Pre-evaluation Comments:
Criteria 2: Scientific Acceptability of Measure Properties (including all 2a, 2b, and 2c)

2a1. Reliability – Specifications

Comments:

- Reliability statistic has increased from 0.88 to 0.98 since the last evaluation. This increase is likely the result of performance on the measure being topped out and perhaps should be considered for Reserve Status.
- Specifications seemed straightforward
- Reliability data elements are well defined.

2a2. Reliability – Testing

Comments:

- No
- No concerns
- Reliability is high and increased from prior measurements up to 98% at the present time. Only concern is it is so high with little gap that this could be considered for reserve status.

2b1. Validity – Testing

Comments:

- Yes because the developer was unable to correlate the measure to an outcome. If we cannot measure a direct correlation between this measure and improved completed mammography rates, then why perform the measure? A couple of our goals should be to streamline and harmonize measures in the American value-based health system and to reduce administrative burden on physicians and providers.
- No strong construct validity, used face validity. I agree with the NQF staff assessment
- Face validity testing conducted in the past; no concerns.

2b4-7. Threats to Validity

Comments:

- For the current submission, the developer performed construct validity. It was unable to correlate the measure to an outcome, so used two process measures, as noted at Item 22. The developer did not find a relationship between the performance rates of the measures; Pearson's correlation coefficients all hover around 0, which is no relationship; of note, both NQF 508 and NQF 509 are nearly "topped out." The developer maintains the measure has high face validity. The developer's assessment of meaningful differences in performance demonstrate there is "no statistically significant difference in measure rates between the top and bottom quartile" and "minor statistically significant and clinically meaningful differences in performance." INSUFFICIENT rating for validity.
- Tables 4-7 on page 58 the standard deviations seem large for claims data. It looks like data are skewed with minimum values of 0.5 % and 0.1% compared to medians of 100%. Are the minimums real or reporting error? Also, the percent missing claims is 8%, which is high. Given the large n of claims it is possible to overpower the analysis

2b2-3. Other Threats to Validity
2b2. Exclusions

2b3. Risk Adjustment

Comments:

- No risk adjustment method was used.
- I didn't see any risk adjustment or exclusions, but this a process measure and not an outcome measure
- Exclusion of those with short life expectancy/not needing mammography is a small group and defensible. No risk adjustment. No other threats to validity identified.

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.

- Data to report the measure are gathered from claims and registries.
- The data are generated or collected by and used by healthcare personnel during the provision of care.
- All data elements are in defined fields in electronic health records (EHRs).

Questions for the Committee:

- None

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

Committee Pre-evaluation Comments:
Criteria 3: Feasibility

3. Feasibility

Comments:

- No concerns--high feasibility rating.
- Appears to be feasible
- Required data elements routinely generated and collected during care delivery. Use of claims data and registry data raises no concerns.

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

- The measure is currently reported in Merit-based Incentives Payment System (MIPS) and has been included in the Physician Quality Reporting System (PQRS) since 2009.

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

The developer reported the following:

- MIPS provides benchmarks annually and MIPS reporters receive QRUR/MIPS reports.
- Quarterly feedback reports are provided to QCDR users that report this measure. ACR staff are available to assist with the interpretation of the measure.
- Feedback is obtained through ACR members, the CMS quality help desk, and CMS contractor QMMS.
- Feedback is considered during the annual measure specification update process with CMS. The ACR Metrics Committee also review the feedback for annual updates.

Additional Feedback: None

***Questions for the Committee**:*

- Can the performance results be used to further the goal of high-quality, efficient healthcare?
- 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

The developer reported the following based on PQRS data:

Year	Average Performance Rate
2010	N/A
2011	68.5 %
2012	74.6 %

2013 81.6%
2015 88.0%
2016 94.4%
2017 96.4%
2018 97.9%

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 The developer did not report any unexpected findings.

Potential harms

- The developer did not report any potential harms.

Additional Feedback:

- None

***Questions for the Committee**:*

- 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: ☐ **High** ☐ **Moderate** ☒ **Low** ☐ **Insufficient**

RATIONALE:

- The measure appears to be topped out, so low usability for audiences (e.g., consumers, purchasers, providers, policymakers).

Committee Pre-evaluation Comments:
Criteria 4: Usability and Use

4a1. Use – Accountability and Transparency

Comments:

- The measure is currently being used in MIPS with transparent data reporting. However, The measure appears to be topped out, so low usability for audiences (e.g., consumers, purchasers, providers, policymakers).
- In use
- Used by CMS nationally. Given the small gaps identified, the resources might be better applied to other quality issues.

4b1. Usability – Improvement

Comments:

- No unintended consequences foreseen. However, the measure appears to be topped out, so low usability for audiences (e.g., consumers, purchasers, providers, policymakers).
- May be some benefit to continue to measure even though it is topped out. subgroup analysis, which wasn't provided may be useful to determine if there are opportunities for improvement.
- Only concern is with potential for over-use either having too frequent mammograms or into older age/short life expectancy women where the benefits might not outweigh the harms. This seems quite

small and perhaps theoretical. Finally, the usability is limited due to the small gap and therefore the measure can be moved to reserve status.

Criterion 5: Related and Competing Measures

Related or competing measures

- The developer identified NQF 2372: Breast Cancer Screening (health plan level) as a related measure.

Harmonization

- Developer states that the measures are harmonized to the extent possible.

Committee Pre-evaluation Comments: Criterion 5: Related and Competing Measures

5. Related and Competing

Comments:

- Given that the measure is basically "topped-out," and one of our goals is to harmonize measures to help reduce reporting burden on physicians and providers, should we focus more on NQF 2372: Breast Cancer Screening (health plan level) since it's a related measure?
- NQF 2372 measures mammography delivered to patients and is within the age group 50-74.

Public and Member Comments

Comments and Member Support/Non-Support Submitted as of: June 30, 2020

- No NQF Members have submitted support/non-support choices as of this date.

Developer Submission

1. Evidence and Performance Gap – 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

NQF_evidence_attachment_0509.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.

No

1a. Evidence (subcriterion 1a)

Measure Number (*if previously endorsed*)**:** 0509

Measure Title: Diagnostic Imaging: Reminder System for Screening Mammograms

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: 4/16/2020

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.

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

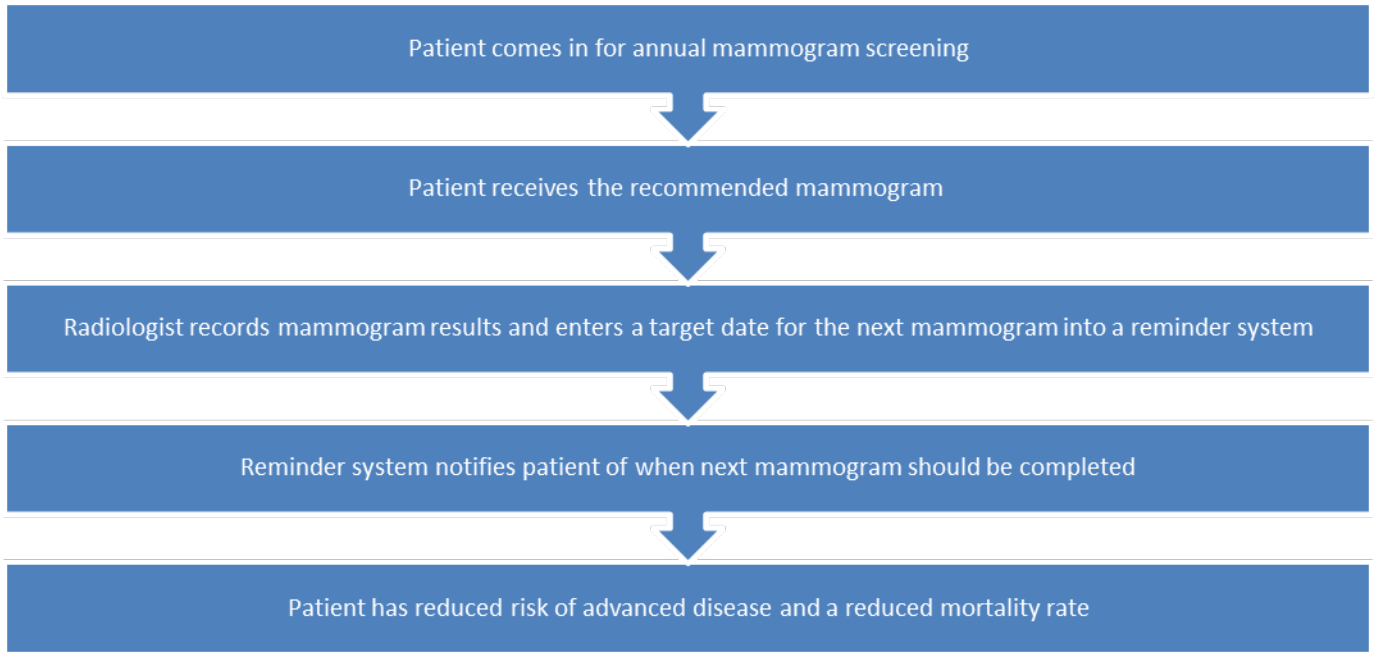
☐ Process:

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

☒ Structure: The system in place to remind patients to come in for mammograms.

☐ 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.



By entering patient information into a reminder system, the patient is more likely to remember to come in for their next screening. Annual screenings ensure that most breast cancers can be detected earlier, which leads to better outcomes for the patient.

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.)

This measure is not derived from the patient report.

****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

Source of Systematic Review: • Title • Author • Date • Citation, including page number • URL	Title: **A 4-year randomized trial comparing three outreach interventions to promote screening mammograms** Authors: **Roger Luckmann, Mary E Costanza, Mary Jo White, Christine F Frisard, Milagros Rosal, Susan Sama, Michelle R Landry, and Robert Yood** Date: **May 23, 2018** Citation: **Roger Luckmann, Mary E Costanza, Mary Jo White, Christine F Frisard, Milagros Rosal, Susan Sama, Michelle R Landry, Robert Yood** **Transl Behav Med. 2019 Apr; 9(2): 328–335. Published online 2018 May 23. doi: 10.1093/tbm/iby031** **https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6610174/**
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.	A simple reminder call can increase screening mammogram adherence even when baseline adherence is high. Some more complex behavioral interventions delivered by mail and phone as in this study may be less effective, due to limited participation of patients, a focus on ambivalence, lack of follow-up, and other factors.
Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	**High:** The evidence is based on random controlled trials and observational studies with a high volume of cases and low to no inconsistencies in the data

Provide all other grades and definitions from the evidence grading system	**High** – random controlled trials/cohort studies with no publication bias and very low to no inconsistencies in the data. **Moderate** – can be RCT/cohorts but with some inconsistencies **Low** – observational studies; inconsistencies in the data and low applicability; limitations in the detailed design and execution **Very low** – many inconsistencies in the data, limitations in detailed design and execution, publication bias, indirectness (PICO and applicability) GRADE grading scale used to evaluate this evidence: https://www.ncbi.nlm.nih.gov/books/NBK138585/
Grade assigned to the **recommendation** with definition of the grade	**Strong –** • (Patients) Most people in this situation would want the recommended course of action and only a small proportion would not •(Clinicians): Most patients should receive the recommended course of action •(Policy makers): The recommendation can be adapted as a policy in most situations
Provide all other grades and definitions from the recommendation grading system	**Strong –** • (Patients) Most people in this situation would want the recommended course of action and only a small proportion would not •(Clinicians): Most patients should receive the recommended course of action •(Policy makers): The recommendation can be adapted as a policy in most situations **Conditional/Weak:** • Patients: The majority of people in this situation would want the recommended course of action, but many would not • Clinicians: Be more prepared to help patients to make a decision that is consistent with their own values/decision aids and shared decision making • Policy makers: There is a need for substantial debate and involvement of stakeholders GRADE grading scale used: https://www.ncbi.nlm.nih.gov/books/NBK138585/

Body of evidence:

- Quantity – how many studies?
- Quality – what type of studies?

This RCT was implemented from 2010 to 2014 at the Fallon Clinic [later renamed Reliant Medical Group (RMG)] which serves Worcester County (population 785,000) in Massachusetts. During the study period, 95 primary care providers (PCPs) served adult RMG patients.

Characteristics of the subjects ever active in the study (≥18 months with Fallon Health and with RMG PCP) by arm and selected characteristics

Characteristics	Counseling call (*n* = 10,054) *n* (%)	Reminder call (*n* = 10,043) *n* (%)	Letter only (*n* = 10,063) *n* (%)	Total (*n* = 30,160) *n* (%)
Age				
40–49	2,246 (22.3)	2,361 (23.5)	2,244 (22.3)	6,851 (22.7)
50–59	2,511 (25.0)	2,529 (25.2)	2,484 (24.7)	7,524 (24.9)
60–69	1,963 (19.5)	1,967 (19.6)	1,959 (19.5)	5,889 (19.5)
70–74	1,025 (10.2)	965 (9.6)	1,047 (10.4)	3,037 (10.1)
75–84	2,309 (23.0)	2,221 (22.1)	2,329 (23.1)	6,859 (22.7)
Race/ethnicity				
White	7,355 (73.2)	7,340 (73.1)	7,416 (73.7)	22,111 (73.3)
Black or African American	176 (1.8)	212 (2.1)	200 (2)	588 (1.9)
Asian	175 (1.7)	152 (1.5)	164 (1.6)	491 (1.6)
American Indian/Alaskan native	91 (0.9)	82 (0.8)	82 (0.8)	255 (0.8)
Native Hawaiian/other Pacific Islander	3 (0.0)	8 (0.1)	6 (0.1)	17 (0.1)
Other	31 (0.3)	36 (0.4)	30 (0.3)	97 (0.3)
Prefer not to answer or unknown	2,223 (22.1)	2,213 (22)	2,165 (21.5)	6,601 (21.9)
Insurance type				
FH Commercial	5,931 (59.0)	6,035 (60.1)	6,000 (59.6)	17,966 (59.6)
FH Medicare	3,613 (35.9)	3,460 (34.5)	3,600 (35.8)	10,673 (35.4)
FH Medicaid	510 (5.1)	548 (5.5)	463 (4.6)	1,521 (5.0)
Ever smoked				
Yes	3,237 (32.2)	3,286 (32.7)	3,168 (31.5)	9,691 (32.1)
No	3,489 (34.7)	3,465 (34.5)	3,564 (35.4)	10,518 (34.9)
Unknown	3,328 (33.1)	3,292 (32.8)	3,331 (33.1)	9,951 (33.0)
Had a mammogram in 24 months prior to study entry				
No	3,076 (30.6)	3,043 (30.3)	3,047 (30.3)	9,166 (30.4)
Yes	6,978 (69.4)	7,000 (69.7)	7,016 (69.7)	20,994 (69.6)

Estimates of benefit and consistency across studies

A key objective of this study was to compare the effectiveness of theory-based, scripted, and computer-supported telephone counseling to a reminder call for promoting mammography when the interventions are delivered repeatedly in a population with a high mammography rate. It was found that reminder calls were more effective than counseling calls and letters only in all age groups from Years 2–4 of the study, although not all differences were statistically significant. At the end of the study, the 40–49-year-old group had the largest difference between the reminder calls and letter only groups, and those aged 50–74 had the least difference. The relative decrease in non-adherence during the study ranged from 16.6% to 21.1% across age groups in the reminder calls group. The lower adherence in the

	40–49 and 75–84-year-old groups was expected, given that some authorities presented screening mammography for these age groups as an option and not as a recommended service. The duration of time in the study was strongly associated with mammography adherence in all types of client reminders. This suggests that for women in the study for 2 years or longer, mammography rates increased over time. Other studies have demonstrated that telephone reminders or counseling can increase repeated mammogram screenings over a period of years. The increase in adherence during Year 1 in all interventions could reflect the change in mailed reminders for the 50- to 74-year-old women to a letter, signed by a woman's primary care provider (PCP) and sent 18 months after their last mammogram. For women between the ages of 40–49 and 75–84, their first experience with any mammography reminder letter came in this study. Other studies have shown that tailored letters, most often from a PCP, are more effective than generic letters. In populations with high mammography utilization, some women may need and respond to a reminder call to adhere to screening recommendations. The result may be a small but possibly cost-effective increase in adherence over a reminder letter. Women aged 40–49 and 75–84 should be engaged in a shared decision-making discussion about breast cancer screening before being encouraged to schedule a mammogram.
What harms were identified?	The main limitations of this study are the socioeconomic and racial/ethnic homogeneity of the population, the high baseline adherence rates, and an inability, due to budget constraints, to interview women who refused counseling or calls. Future studies should include more diverse populations with adherence rates more representative of the general population and should plan to gather data on women refusing counseling or calls
Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	No, evidence has remained consistent that client reminders increase the frequency of mammograms.

☐ 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

Source of Systematic Review: • Title • Author • Date • Citation, including page number • URL	Title: **Effectiveness of Interventions to Increase Screening for Breast, Cervical, and Colorectal Cancers: Nine Updated Systematic Reviews for the Guide to Community Preventive Services** Authors: **Susan A. Sabatino, MD, MPH, Briana Lawrence, MPH, Randy Elder, PhD, MEd, Shawna L. Mercer, MSc, PhD, Katherine M. Wilson, PhD, MPH, Barbara DeVinney, PhD, Stephanie Melillo, MPH, Michelle Carvalho, MPH, Stephen Taplin, MD, MPH, Roshan Bastani, PhD, Barbara K. Rimer, DrPH, Sally W. Vernon, PhD, Cathy Lee Melvin, PhD, MPH, Vicky Taylor, BMBS, MPH, Maria Fernandez, PhD, Karen Glanz, PhD, MPH, and the Community Preventive Services Task Force** Date: **April 2012** Citation: **Sabatino SA, Lawrence B, Elder R, et al. Effectiveness of interventions to increase screening for breast, cervical, and colorectal cancers: nine updated systematic reviews for the guide to community preventive services. Am J Prev Med. 2012;43(1):97-118. doi:10.1016/j.amepre.2012.04.009** **https://pubmed.ncbi.nlm.nih.gov/22704754/**
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.	According to Community Guide rules of evidence, there is strong evidence that client reminders are effective in increasing screening for breast and cervical cancers. For provider-directed interventions, audit and feedback have been associated with increased mammography screening.
Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	**Recommended:** Strong evidence to support a recommendation
Provide all other grades and definitions from the evidence grading system	**Insufficient evidence** – not enough evidence to support a recommendation.
Grade assigned to the **recommendation** with definition of the grade	**Strong –** • (Patients) Most people in this situation would want the recommended course of action and only a small proportion would not

	•(Clinicians): Most patients should receive the recommended course of action •(Policy makers): The recommendation can be adapted as a policy in most situations
Provide all other grades and definitions from the recommendation grading system	**Strong –** • (Patients) Most people in this situation would want the recommended course of action and only a small proportion would not •(Clinicians): Most patients should receive the recommended course of action •(Policy makers): The recommendation can be adapted as a policy in most situations **Conditional/Weak:** • Patients: The majority of people in this situation would want the recommended course of action, but many would not • Clinicians: Be more prepared to help patients to make a decision that is consistent with their own values/decision aids and shared decision making • Policy makers: There is a need for substantial debate and involvement of stakeholders GRADE grading scale used: https://www.ncbi.nlm.nih.gov/books/NBK138585/
Body of evidence: • Quantity – how many studies? • Quality – what type of studies?	The original review of client reminders found strong evidence of effectiveness based on a median increase of 14.0 percentage points in recent mammography (19 studies) and three additional studies demonstrating an increase in repeat mammography. In the update, six additional studies were included. All had greatest design suitability except for one with a least suitable design. Exclusion of this study did not change overall conclusions.
Estimates of benefit and consistency across studies	Outcomes for update studies of breast cancer screening promotion were determined via self-report, medical record review, administrative records, or screening program attendance. Interventions included both textual and telephone reminders, which included automated interactive voice response reminders (AIVR) by phone as well as tailored interventions and enhanced interventions (as in the original

	review, defined as including follow-up reminders, additional text, discussion, or appointment scheduling assistance). Studies included reminders delivered by clinical practices or organizations, screening programs or registries, or other sources. Interventions were conducted in the U.S. and Norway. Participants included white, African-American, and Hispanic participants. Some studies included groups of unspecified race and others did not report race or ethnicity. Individuals with low and urban or mixed urban/rural populations also were included. Several studies did not report this information. Of four update studies providing information about absolute change in mammography use, two provided information about recent screening only, defined as completion of the most recent mammogram within a specified interval; one provided information about repeat mammography only, defined as examining two or more consecutive, on-time mammograms; and one provided information about both. The only phone intervention among these four studies was the AIVR study. When studies from both reviews were combined to examine differences by recent versus repeat screening use, the median increase for recent use was 12.3 percentage points and for repeat mammography was 6.0 percentage points. Findings from the original review also suggested that unenhanced, printed reminders have smaller effects than enhanced or telephone reminders (median 3.6 percentage points across 12 studies vs 18.5 percentage points across 13 studies, respectively). This conclusion was supported by all nine intra-study comparisons. One study with separate arms for unenhanced and enhanced client reminders was incorporated, and the findings reaffirmed that enhanced or telephone reminders may have a greater effect (15.5 percentage points vs 4.5 percentage points). The team also examined the incremental effect of client reminders beyond the effect of other intervention components common to two or more study arms. One study in the update, six studies in the original review, and two studies from the review of multicomponent interventions enabled this type of comparison. Across all nine studies, the overall median ~~incremental effect was 5.~~0 percentage points.

What harms were identified?	No reports of other positive or negative effects of interventions on use of other healthcare services, health behaviors, or informed decision making were found during this review. For client reminders, barriers may include limited infrastructure and staffing and/or computer support to identify patients due for screening and deliver reminders efficiently. Costs of generating and delivering reminders may be a substantial barrier, as well as barriers related to tailoring.
Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	No, evidence has remained consistent that client reminders increase the frequency of mammograms.

☒ 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

Source of Systematic Review: • Title • Author • Date • Citation, including page number • URL	Title: **Updated recommendations for client- and provider-oriented interventions to increase breast, cervical, and colorectal cancer screening** Authors: **Community Preventative Services Task Force** Date: **2012** Citation: **Community Preventive Services Task Force. Updated recommendations for client- and provider-oriented interventions to increase breast, cervical, and colorectal cancer screening. Am J Prev Med. 2012;43(1):92-96. doi:10.1016/j.ampre.2012.04.008.** URL: **http://www.thecommunityguide.org/cancer/screening/client-oriented/reminders.html**
Quote the guideline or recommendation verbatim about the	The Community Preventive Services Task Force recommends the use of client reminders to increase screening for breast and

process, structure or intermediate outcome being measured. If not a guideline, summarize the conclusions from the SR.	cervical cancers on the basis of strong evidence of effectiveness. The Task Force also recommends the use of client reminders to increase colorectal cancer screening with fecal occult blood testing based on strong evidence of effectiveness. Evidence is insufficient, however, to determine effectiveness of client reminders in increasing colorectal cancer screening with other tests (colonoscopy, flexible sigmoidoscopy), because of inconsistent evidence.
Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	The Community Preventive Services Task Force (Task Force) uses the terms below to describe its findings. **Grade: Recommended** The systematic review of available studies provides strong or sufficient evidence that the intervention is effective.
Provide all other grades and definitions from the evidence grading system	The categories of "strong" and "sufficient" evidence reflect the Task Force's degree of confidence that an intervention has beneficial effects. They do not directly relate to the expected magnitude of benefits. The categorization is based on several factors, such as study design, number of studies, and consistency of the effect across studies
Grade assigned to the **recommendation** with definition of the grade	**Strong –** • (Patients) Most people in this situation would want the recommended course of action and only a small proportion would not •(Clinicians): Most patients should receive the recommended course of action •(Policy makers): The recommendation can be adapted as a policy in most situations
Provide all other grades and definitions from the recommendation grading system	**Strong –** • (Patients) Most people in this situation would want the recommended course of action and only a small proportion would not •(Clinicians): Most patients should receive the recommended course of action •(Policy makers): The recommendation can be adapted as a policy in most situations **Conditional/Weak:** • Patients: The majority of people in this situation

	would want the recommended course of action, but many would not • Clinicians: Be more prepared to help patients to make a decision that is consistent with their own values/decision aids and shared decision making • Policy makers: There is a need for substantial debate and involvement of stakeholders
Body of evidence: • Quantity – how many studies? • Quality – what type of studies?	Nineteen studies qualified for the systematic review. All of the study designs were based on a randomized trial (individual). Study locations varied from all areas of the US.
Estimates of benefit and consistency across studies	
What harms were identified?	
Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	

☐ 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

Source of Systematic Review: • **Title** • **Author** • **Date** • **Citation, including page number** • **URL**	**Title: Screening for Breast Cancer: A Systematic Review to Update the 2009 U.S. Preventive Services Task Force Recommendation** **Author: Nelson HD, Cantor A, Humphrey L, et al** **Date: January 2016** **Citation: Nelson HD, Cantor A, Humphrey L, et al. Screening for Breast Cancer: A Systematic Review to Update the 2009 U.S. Preventive Services Task Force Recommendation**

	[Internet]. Rockville (MD): Agency for Healthcare Research and Quality (US); 2016 Jan. (Evidence Syntheses, No. 124.) Available from: http://www.ncbi.nlm.nih.gov/books/NBK343819/
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.	**Women aged 50 to 74 years** The USPSTF recommends biennial screening mammography for women aged 50 to 74 years. These recommendations apply to asymptomatic women aged 40 years or older who do not have preexisting breast cancer or a previously diagnosed high-risk breast lesion and who are not at high risk for breast cancer because of a known underlying genetic mutation (such as a *BRCA1* or *BRCA2* gene mutation or other familial breast cancer syndrome) or a history of chest radiation at a young age.
Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	**High -** random controlled trials/cohort studies with no publication bias and very low to no inconsistencies in the data.
Provide all other grades and definitions from the evidence grading system	**High** – random controlled trials/cohort studies with no publication bias and very low to no inconsistencies in the data. **Moderate** – can be RCT/cohorts but with some inconsistencies **Low** – observational studies; inconsistencies in the data and low applicability; limitations in the detailed design and execution **Very low** – many inconsistencies in the data, limitations in detailed design and execution, publication bias, indirectness (PICO and applicability) GRADE grading scale used to evaluate this evidence: https://www.ncbi.nlm.nih.gov/books/NBK138585/
Grade assigned to the **recommendation** with definition of the grade	**The USPSTF recommends biennial screening mammography for women aged 50 to 74 years.** This recommendation is assigned a 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. Suggestions for the practice are to offer or provide this service. **Women aged 40 to 49 years**

	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. 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. • For women who are at average risk for breast cancer, most of the benefit of mammography results from biennial screening during ages 50 to 74 years. Of all of the age groups, women aged 60 to 69 years are most likely to avoid breast cancer death through mammography screening. While screening mammography in women aged 40 to 49 years may reduce the risk for breast cancer death, the number of deaths averted is smaller than that in older women and the number of false-positive results and unnecessary biopsies is larger. The balance of benefits and harms is likely to improve as women move from their early to late 40s. • In addition to false-positive results and unnecessary biopsies, all women undergoing regular screening mammography are at risk for the diagnosis and treatment of noninvasive and invasive breast cancer that would otherwise not have become a threat to their health, or even apparent, during their lifetime (known as "overdiagnosis"). Beginning mammography screening at a younger age and screening more frequently may increase the risk for overdiagnosis and subsequent overtreatment. • Women with a parent, sibling, or child with breast cancer are at higher risk for breast cancer and thus may benefit more than average-risk women from beginning screening in their 40s. Go to the Clinical Considerations section for information on implementation of the C recommendation. 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.

	Women aged 75 years or older Grade I - 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. **Al Women** - 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. **Women with dense breasts** - 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.
Provide all other grades and definitions from the recommendation grading system	**Grade A -** The USPSTF recommends the service. There is high certainty that the net benefit is substantial. Offer or provide this service. **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. Offer or provide this service. **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. Offer or provide this service for selected patients depending on individual circumstances. **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. Discourage the use of this service. **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.

	http://www.uspreventiveservicestaskforce.org/Page/Document/RecommendationStatementFinal/breast-cancer-screening1
Body of evidence: • Quantity – how many studies? • Quality – what type of studies?	The USPSTF commissioned a series of systematic evidence reviews in support of this recommendation. The first addressed the effectiveness of breast cancer screening in reducing breast cancer–specific and all-cause mortality, as well as the incidence of advanced breast cancer and treatment-related morbidity. It also looked at the harms of breast cancer screening. A second systematic review summarized the evidence about the test performance characteristics of DBT as a primary screening strategy. A third systematic review evaluated the evidence on adjunctive screening in women with increased breast density, including the accuracy and reproducibility of dense breast classification systems and the diagnostic test performance characteristics, benefits, and harms of adjunctive screening in women identified to have dense breasts on an otherwise negative screening mammogram. In total, this guideline referenced 62 studies.
Estimates of benefit and consistency across studies	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 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.
What harms were identified?	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.
Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	This is the most updated guideline. There are no newer studies to change the current frequency of mammograms.

☐ 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

Source of Systematic Review: • **Title** • **Author** • **Date** • **Citation, including page number** • **URL**	**Systematic review #1 (Baron et al 2010):** **http://www.thecommunityguide.org/cancer/screening/provider-oriented/InterventionsIncreaseRecommendationDeliveryScreeningBreastCervicalColorectalCancersHealthcareProvidersSystematicReview_2.pdf** **Systematic review #2 (Baron et al 2008)** **http://www.thecommunityguide.org/cancer/screening/client-oriented/Cancer2008_ClientDirected_Demand.pdf**
Quote the guideline or recommendation verbatim about the process, structure or intermediate	Systematic review #1 (Baron et al 2010)

outcome being measured. If not a guideline, summarize the conclusions from the SR.	This report presents results of systematic reviews of effectiveness, applicability, economic efficiency, barriers to implementation, and other harms or benefits of provider reminder/recall interventions to increase screening for breast, cervical, and colorectal cancers. Evidence in this review of studies published from 1986 through 2004 indicates that reminder/recall systems can effectively increase screening with mammography, Pap, fecal occult blood tests, and flexible sigmoidoscopy. Systematic review #2 (Baron et al 2008) This report presents the results of systematic reviews of effectiveness, applicability, economic efficiency, barriers to implementation, and other harms or benefits of interventions designed to increase screening for breast, cervical, and colorectal cancers by increasing community demand for these services. Evidence from these reviews indicates that screening for breast cancer (mammography) and cervical cancer (Pap test) has been effectively increased by use of client reminders, small media, and one-on-one education.
Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	The systematic reviews identified in this application use a similar evaluation of studies for review *Guide to Community Preventive Services*. Each study is characterized based on both the suitability of study design for assessing effectiveness and the quality of study execution. Study designs are classified using a standard algorithm Greatest - concurrent comparison groups *and* prospective measurement of exposure and outcome Moderate - all retrospective designs *or* multiple pre or post measurements but no concurrent comparison group Least - single pre and post measurements and no concurrent comparison group *or* exposure and outcome measured in a single group at the same point in time
Provide all other grades and definitions from the evidence grading system	**Quality of Execution** Each study is categorized as having good, fair, or limited quality of execution based on the number of limitations noted, studies with 0–1, 2–4, and 5 or more limitations are categorized as having good, fair, and limited execution respectively. Studies with limited execution are not included in bodies of evidence to support recommendations. In general, information on quality of study execution is based only on information in published reports because

	bias could be introduced based on limited availability or variable quality of additional information from the authors and because collecting additional information from the authors may not be feasible. Several principles guided the designation of bodies of evidence of effectiveness as strong, sufficient, or insufficient evidence. Strong or sufficient evidence can be based either on a small number of studies with better execution and more suitable design or a larger number of studies with less suitable design or weaker execution Briss PA, Zaza S, Pappaioanou M, et al. Developing an evidence-based Guide to Community Preventive Services— methods. Am J Prev Med 2000;18(1S):35– 43. http://www.thecommunityguide.org/about/methods-ajpm-developing-guide.pdf
Grade assigned to the **recommendation** with definition of the grade	**The USPSTF recommends biennial screening mammography for women aged 50 to** harms cannot be determined. 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.
Provide all other grades and definitions from the recommendation grading system	**Grade A -** The USPSTF recommends the service. There is high certainty that the net benefit is substantial. Offer or provide this service. **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. http://www.uspreventiveservicestaskforce.org/Page/Document/RecommendationStatementFinal/breast-cancer-screening1
Body of evidence: • Quantity – how many studies? • Quality – what type of studies?	Systematic review #1 (Baron et al 2010) The search for evidence identified 38 studies that reported on using provider reminders to increase recommended screening for breast, cervical, and colorectal cancers. Of these, six were excluded because of their low quality of execution and six more were

	excluded because of the lack of a concurrent comparison group. Of the 26 remaining studies that qualified for review, five had good quality of execution, and 21 studies had fair quality of execution. Of the studies that qualified for review 9 were from randomized control trials and 16 were observational studies. Systematic review #2 (Baron et al 2008) The searches for evidence identified 39 studies of greatest design suitability were identified that reported using client reminders to increase breast cancer screening by mammography. Of these, nine studies were excluded due to limited quality of execution and were excluded because comparison groups received different reminders or reminders of lesser intensity than study groups. Of the 19 remaining studies that qualified for review, had fair quality of execution and two had good quality of execution. Six studies, five classified as good and one as very good met inclusion criteria for cost-effectiveness analysis of client reminders in increasing breast cancer screening by mammography. 2010 Baron et al The qualifying studies examined mammography Pap, and colorectal screening. All measured outcomes (screening tests completed, or screening tests recommended or ordered but not necessarily completed) were ascertained by record review.13 studies for mammography – pertained to the primary outcome of interest, completed screening tests. Mammography screening increased by a median of 10.0% (IQI, 3.0%–19.0) for all screening modalities, but in particular for mammography, the absolute effect of provider reminders on completed screenings appears to have diminished over time. Because background screening rates often were not provided for study populations, the role, if any, of temporal changes in baseline screening rates on these results could not be determined. Evidence in this review of studies published from 1986 through 2004 indicates that reminder/recall systems can effectively increase screening with mammography, Pap, fecal occult blood tests, and flexible sigmoidoscopy. Baron et al 2008 Twenty studies were identified that reported using small media to increase breast cancer screening by mammography.

	One study was excluded due to limited quality of execution. Of 19 qualifying studies, 17 had greatest design suitability, of which three had good quality of execution and 14 had fair quality of execution. Two qualifying studies, one with moderate and one with least suitable study design, had fair quality of execution. Five studies evaluated tailored interventions, twelve evaluated untailored interventions, and two studies included both a tailored and an untailored intervention.
Estimates of benefit and consistency across studies	Systematic review #1 (Baron et al 2010) The original review included 19 studies. This update included an additional 6 studies. Combined evidence from both the original and the updated review showed the following. • Mammography screening: median increase of 14.0 percentage points (interquartile interval [IQI]: 2.0 to 24.0 percentage points; 19 studies with 32 study arms). • Recent mammography screening: median increase of 12.3 percentage points (IQI: 3.0 to 18.9 percentage points; 30 study arms). • Repeat mammography screening: median increase of 6.0 percentage points (IQI 3.0 to 19.1 percentage points; 8 study arms). • Enhanced and telephone reminders showed a greater increase (15.5 percentage points [IQI 7.0 to 29.0 percentage points]; 20 study arms) than written reminders alone (4.5 percentage points [IQI: 1.9 to 14.0 percentage points]; 14 study arms). • When added to other types of interventions, the median incremental effect for client reminders was an increase of 5.0 percentage points (IQI 1.6 to 6.7 percentage points; 12 study arms). Client reminder interventions to increase breast cancer screening should be applicable across a range of settings and populations, provided they are adapted to the target population and delivery context. Systematic Review #2 (Baron et al 2008) According to *Community Guide* methods, there is strong evidence that client reminders increase breast and cervical cancer screening by mammography and Pap test, respectively. These findings should apply across a range of settings and populations. Although evidence also suggests

	that enhancement of simple printed reminders with additional messages or support to clients results in greater effectiveness, particularly for breast cancer screening, it is not yet known whether such enhancement increases effectiveness among women who have never been screened or who may be hard to reach. Overall, the median post-intervention increase in completed mammography was 14.0 percentage points (interquartile interval [IQI]= 2.0, 24.0). The magnitude of this effect and consistent positive results across studies and reminder systems demonstrate the effectiveness of client reminders in increasing breast cancer screening by mammography. 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.
What harms were identified?	No reports of benefits or harms related to the use of provider reminders were found. Potential benefits include increases in the use of other preventive services linked to the reminder system.
Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	

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.

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

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

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

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.

Although screening mammograms can reduce breast cancer mortality by 20-35% in women aged 40 years and older, recent evidence shows that only 72% of women are receiving mammograms based on current guideline recommendations. The use of patient reminders is associated with an increase in screening mammography.Encouraging the implementation of a reminder system could lead to an increase in mammography screening at appropriate intervals.

Any facility that uses less than annual frequency of screening, greatly increases the importance of attendance at each scheduled screening. Even with annual screening recommendations screening does not always occur biennially. This demonstrates the importance of systematic reminders and active patient outreach.

The purpose of screening is to minimize interval or false negative cancers, as these are failures of the screening process. The 2011 article by Bennett, Sellars and Moss (Ref 1)and an earlier work by Woodman,Threlfall and Boggis (ref 2) examine the effect of interval cancer rates (false negative cancers) by time since screen out to three years in the United Kingdom´s triennial screening program. The Interval cancer rates (false negative cases) increase over time (ref 1,2) and begin to approach incidence rates by the third year (ref 2). Thus screening at greater than 2 year intervals will likely have poor overall outcomes in reducing breast cancer mortality. These papers may also underestimate the rate of interval cancers (ref 1) so the actual rates may be higher.

Efforts to ensure regular screening are therefore necessary to eliminate any screening interval beyond 2 years.

1. Bennett RL, Sellars SJ, Moss SM. Interval cancers in the NHS breast cancer screening programme in England, Wales and Northern Ireland. British Journal of Cancer. 2011. 104: 571-577.

2. Woodman CBJ, Threlfall AG, Boggis CR et al. Is the three year breast screening interval too long? Occurrence of interval cancers in NHS breast screening programme's north western region. BMJ. 1995. 310:224-6

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.*

This measure was included in the CMS Physician Quality Reporting System as measure #225 Reminder System for Screening Mammograms from 2009 until 2016, and in the CMS Merit-based Incentives Payment System from 2017 until now. There is a gap in care as shown by this data; between 2012 and 2014 15.0 % of patients reported on did not meet the measure.

Scores on this measure for 2012-2014 (calculated using data from CMS):

N=47,866 physicians with at least 10 patients had a non-zero reporting rate. Across these physicians, 15.0% of physicians did not meet the measure (100% - 85.0% who met the measure). Across physicians with at least 10 patients and a performance rate greater than zero for the 3-year period 2012-2014, mean performance rate= 85.0%.

The performance rate quartiles for the same period 2012-2014 for physicians with at least 10 patients and performance rate >0 were as follows:

Scores on this measure is N= 47,866

25th percentile: 91.15%

50th percentile: 100%

75th percentile: 100%

Exception Rate: This measure is not specified with exceptions. See attached performance data.

The performance rate for the three year period was calculated as the count of reported instances where performance was met (numerator=6,423,710) divided by the total number of reported instances (7,554,604). Performance rate was also calculated in this way for each year (2012-2014) with the following results:

2012-2014 85.0%

2012 79.4%

2013 86.0%

2014 87.6%

Among 47866 total physicians included in the analysis, 45380 submitted data by claims, and 2486 submitted data by registry (reporting_method). For our analyses we used the combined total of 47866 for both claims and registry reported cases.

Scores on this measure for 2015-2018 (calculated using data from CMS):

N= 79,450 physicians with at least 10 patients had a non-zero reporting rate. Across these physicians, 7.4% of physicians did not meet the measure. Across physicians with at least 10 patients and a performance rate greater than zero for the 4-year period 2015-2018, mean performance rate= 95.69%.

The performance rate quartiles for the same period 2015-2018 for physicians with at least 10 patients and performance rate >0 were as follows:

25th percentile: 99.5%

50th percentile: 100%

75th percentile: 100%

• The performance rate for the four year period was calculated as the count of reported instances where performance was met (numerator= 65,874,442) divided by the total number of reported instances (68,844,412). Performance rate was also calculated in this way for each year (2015-2018) with the following results:

2015-2018 95.7%

2015 88.0%

2016 94.4%

2017 96.4%

2018 95.6%

•Among 79,450 total physicians included in the analysis, 69,240 submitted data by claims, and 10,210 submitted data by registry (either QCDR or qualified registry). For our analyses we used the combined total of 79,450 for both claims and registry reported cases.

Rationale for Performance Calculations

•Medicare claims data with information on reporting measure #225 from years 2012-2018 was used for performance calculation and analyses.

•For each year, if the patient's eligible (pts_eligible) for a particular physician (npi) was greater or equal to 10, the physician was included in the analysis

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.

There is sufficient performance data.

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.*

Many American women do not receive mammograms at recommended intervals, as illustrated by 2010 data from the National Health Interview Survey (NHIS) which found that only 72% of women reported receiving a mammogram within the recommended two-year interval. Additional factors found to reduce the likelihood for a woman to receive a mammogram include Asian race, low education status, and recent immigrant status. Low mammography use was also noted for women who reported having no regular source of medical care or having no medical insurance.

Centers for Disease Control and Prevention (CDC). Cancer screening—United States, 2010. MMWR 2012;61(3):41-45. http://www.cdc.gov/mmwr/preview/mmwrhtml/mm6103a1.htm. Accessed 2/3/2014.

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

The American College of Radiology advocates that performance measure data should, where possible, be stratified by race, ethnicity, and primary language to assess disparities and initiate subsequent quality improvement activities addressing identified disparities, consistent with recent national efforts to standardize the collection of race and ethnicity data.

A 2008 NQF report endorsed 45 practices including stratification by the aforementioned variables (1). A 2009 IOM report "recommends collection of the existing Office of Management and Budget (OMB) race and Hispanic ethnicity categories as well as more fine-grained categories of ethnicity (referred to as granular ethnicity and based on one´s ancestry) and language need (a rating of spoken English language proficiency of less than very well and one´s preferred language of health-related encounters)." (2)

1. National Quality Forum Issue Brief (no. 10) Closing the Disparities Gap in Healthcare Quality with Performance Measurement and Public Reporting. Washington, DC: NWF, August 2008.

2. Race, Ethnicity, and Language Data: Standardization for Health Care Quality Improvement. March 2010. AHRQ Publication No. 10-0058-EF. Agency for Healthcare Research and Quality, Rockville, MD. Available at:

http://www.ahrq.gov/research/iomracereport. Accessed May 25, 2010.

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):*

Care Coordination, Screening

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

Populations at Risk

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.)*

https://www.acr.org/-/media/ACR/NOINDEX/Measures/2020_Measure_225_MIPSCQM.pdf

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)*

No data dictionary **Attachment:**

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.

No

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.

There are no significant changes since last endorsement.

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).

Patients whose information is entered into a reminder system with a target due date for the next mammogram

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).

Numerator Note:

The reminder system should be linked to a process for notifying patients when their next mammogram is due and should include the following elements at a minimum: patient identifier, patient contact information, dates(s) of prior screening mammogram(s) (if known), and the target due date for the next mammogram. Use of the reminder system is not required to be documented within the final report to meet performance for this measure.

Performance Met: Patient information entered into a reminder system with a target due date for the next mammogram (7025F)

Performance Not Met: Patient Information not entered into a reminder system, reason not otherwise specified (7025F with 8P)

S.6. Denominator Statement *(Brief, narrative description of the target population being measured)*

All patients undergoing a screening mammogram

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).

Denominator Criteria (Eligible Cases):

All patients, regardless of age

AND

Diagnosis for mammogram screening (ICD-10-CM): Z12.31

AND

Patient procedure during the performance period (CPT or HCPCS): 77067

S.8. Denominator Exclusions *(Brief narrative description of exclusions from the target population)*

Documentation of medical reason(s) for not entering patient information into a reminder system [(eg, further screening mammograms are not indicated, such as patients with a limited life expectancy, other medical reason(s)]

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.)*

Documentation of medical reason(s) for not entering patient information into a reminder system (e.g., further screening mammograms are not indicated, such as patients with a limited life expectancy, other medical reason(s) (7025F with 1P)

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.)*

We encourage the results of this measure to be stratified by race, ethnicity, sex, and payer.

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.)*

To calculate performance rates:

1) Find the patients who meet the initial patient population (ie, the general group of patients that the performance measure is designed to address).

2) From the patients within the initial patient population criteria, find the patients who qualify for the denominator (ie, the specific group of patients for inclusion in a specific performance measure based on defined criteria). Note: in some cases the initial patient population and denominator are identical.

3) From the patients within the denominator, find the patients who qualify for the Numerator (ie, the group of patients in the denominator for whom a process or outcome of care occurs). Validate that the number of patients in the numerator is less than or equal to the number of patients in the denominator

If the patient does not meet the numerator, this case represents a quality failure.

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.

This measure is not based on a sample or survey.

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, Registry 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.

We´re using data submitted to CMS through claims and registries for the Merit-based Incentives Payment Program.

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)*

Clinician : Individual

S.21. Care Setting *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*

Inpatient/Hospital, 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.)*

This is not a composite measure.

2. Validity – See attached Measure Testing Submission Form

NQF_Testing_Attachment_2019_225-637226353068492536.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.

No

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

Measure Testing (subcriteria 2a2, 2b1-2b6)

Measure Number (*if previously endorsed*)**: 0**509

Measure Title: Diagnostic Imaging: Reminder System for Screening Mammograms

Date of Submission: January 6, 2020

Type of Measure:

☐ Outcome (*including PRO-PM*)	☐ Composite – ***STOP – use composite testing form***
☐ Intermediate Clinical Outcome	☐ Cost/resource
■ Process *(including Appropriate Use)*	☐ Efficiency
☐ Structure	

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*).

Data was obtained from CMS for all payers.

The American College of Radiology (ACR) completed measure testing using both claims and registry data for the Medicare, Medicaid and commercial payer populations. The data sources were obtained directly from CMS for all populations.

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

2012-2014

January 1, 2015 - December 31, 2018

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*)

The numbers of physicians were 47,866 physicians
Among these physicians 45,380 were claims and 2,486 were from registry

- The data collection period was 2012-2014
- Data abstraction was performed in 2015

The testing sample is comprised of all NPIs that submitted data to CMS for this measure. The sample consisted **of 79,450 physicians**.

Table 1. Number of providers that submitted data for this measure

	# of NPIs
All- Claims, QCDR, and Registry	
All 4 Years	**79450**
2015	**21952**
2016	**25556**
2017	**18292**
2018	**13650**
Claims	
All 4 Years	**69240**
2015	**18724**
2016	**20570**
2017	**17298**
2018	**12648**
QCDR	
All 4 Years	**450**
2015	**316**
2016	**134**

Registry	
All 4 Years	**9760**
2015	**2912**
2016	**4852**
2017	**994**
2018	**1002**

1.6. How many and which <u>patients</u> 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*)

- Number of patients eligible were 14,515,814 (avg. per NPI is 303.26)
- Number of patients reported were 7,554,604 (avg. per NPI is 157.83)

Update 2020: The eligible population for this measure (i.e. the denominator) includes all patients, regardless of age, undergoing a screening mammogram. Patients with documented medical reason(s) for not being placed into a reminder system (e.g., further screening mammograms are not indicated, such as patients with a limited life expectancy, other medical reason(s) are removed from the eligible population. The measured entities are not limited to Medicare Part B patients, with the testing sample originating from a broad-swath of US locations (e.g., small and rural locations, urban, ambulatory). The following testing analysis used the number of patients reported that were eligible for the measure, irrelevant of health care plan provider.

Table 2. Eligible Patients and Reported Patients

	# of Patients Eligible	# of Patients Reported
All- Claims, QCDR, and Registry		
All 4 Years	**74792218**	**68844412**
2015	**8815174**	**7960504**
2016	**14070952**	**12904940**
2017	**20294820**	**18484274**
2018	**31611272**	**29494694**
Claims		
All 4 Years	**24685210**	**20210596**
2015	**6576870**	**5742550**
2016	**7585340**	**6642078**
2017	**5521408**	**4310084**
2018	**5001592**	**3515884**
QCDR		
All 2 Years	**4107514**	**3905360**
2015	**445554**	**442150**
2016	**3661960**	**3463210**
Registry		
All 4 Years	**45999494**	**44728456**
2015	**1792750**	**1775804**
2016	**2823652**	**2799652**
2017	**14773412**	**14174190**
2018	**26609680**	**25978810**

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.

There are no differences in the data used for testing.

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.

No social risk factors for this measure.

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*)

ACR performed a signal-to-noise ratio (SNR) analysis test on the performance data for reliability. In SNR analysis, reliability is the measure of confidence in differentiating performance between physicians or other providers. The signal is the variability in measured performance that can be explained by real differences in physician performance and the noise is the total variability in measured performance.

A reliability score equal to zero implies that all the variability in a measure is attributable to measurement error. A reliability score equal to one implies that all the variability is attributable to real differences in physician performance. A reliability score of 0.70 is generally considered the minimum threshold for reliability and 0.80 is generally considered very good reliability.

SNR reliability testing is performed using the Beta-Binomial Model, which assumes that physicians' performance scores are a binomial random variable conditional on the physicians' true value derived from the beta distribution. The beta distribution is usually defined by two parameters, alpha and beta. Alpha and beta are considered intermediate calculations used to establish the variance estimates.

The steps were taken to estimate Alpha and Beta:
1) Build a data file of the proper form for physician-to-physician variance estimation.
2) Use the Beta in SAS macro to estimate the physician-to-physician variance.
3) Use the physician-to-physician variance estimate and the physician-specific information to calculate the physician specific reliability scores.

ACR testing protocol followed the convention of estimating reliability at two points: 1) at a minimum number of qualities reporting events per physician and 2) at the average number of quality reporting events per physician. The minimum threshold of events was set at 10. Limiting the reliability analysis to physicians with a minimum number of events reduces bias introduced by the inclusion of physicians without a significant numbers of events.

CMS physician-level claims, registry, and QCDR data was extracted for the relevant physician-level information.

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*)

Physician to Physician variation stats- 225

Label	*Estimate*	*StandardError*	*tValue*	*Probt*	*Alpha*	*Lower*	*Upper*
mu	0.2867	0.00168	170.77	<.0001	0.05	0.2834	0.2899
alpha	0.0693	0.0006	115.34	<.0001	0.05	0.06813	0.0705
beta	0.1725	0.00147	117.15	<.0001	0.05	0.1696	0.1754

Summary of PQRS Reliability Score Stats by Year (2012 - 2014)

Year	*Number of Providers*	*Reliability p25*	*Reliability median*	*Reliability p75*	*Reliability mean*	*Reliability LCLM*	*Reliability UCLM*
2012	19955	1	1	1	0.87513	0.87134	0.87892
2013	18427	0.81469	1	1	0.85736	0.85371	0.86101
2014	9484	0.98538	0.99698	0.99977	0.98146	0.98057	0.98234
All	47866	0.96641	1	1	0.88936	0.88719	0.89152

Using the parameter estimates from the beta-binomial model, we computed reliability scores for each performance year. Please see **Table 3** for the results.

Table 3. Reliability Score Statistics by Year by Provider (claims and registry)

Year	*Number of Providers*	*25th percentile*	*Reliability median*	*75th percentile*	*Reliability mean*	*Lower Confidence Limit (minimum)*	*Upper Confidence Limit (maximum)*
2015	44256	.98402	.99692	1.00000	.97984	.97939	.98030
2016	52048	.99158	.99875	1.00000	.98876	.98850	.98902
2017	39860	.98574	.99742	.99996	.98349	.98313	.98384
2018	30648	.98340	.99691	.99999	.97791	.97729	.97854
ALL	166812	.98713	.99775	1	.98314	.98294	.98335

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?*)

The mean (CI), P25, median, P75 of the reliability score results are shown in the above table for all 3 years as well as by each year. Our mean (CI) reliability is 0.88936 (0.88719, 0.89152). A reliability of 0.80 is considered very good reliability. So according to the reliability testing analysis, the results demonstrated very good reliability.

This measure has proven to be consistent and dependable. Using the total number of providers from 2015-2018 (166,812) the performance data was analyzed and produced a mean reliability is of .98, which is higher than testing completed in 2015. Over time, the measure has continued to produce similar results and performance has increased.

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

X Empirical validity testing

☐ **Systematic assessment of face validity of <u>performance measure score</u> 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)*

We are not able to establish strong construct validity for this measure, but maintain that the measure has high face validity. For construct validity, we are unable to correlate NQF #509 to an outcome like cancer. However, we describe efforts to demonstrate construct validity using computed performance scores of measures with similar effects in Tables 1 and 2 below. Table 3 provides performance rate comparisons by facility types; this analysis was conducted to determine if there was potential disparities.

NQF #509, compared against NQF #508: *Inappropriate Use of "Probably Benign" Assessment Category in Screening Mammograms*, and ACRad5: *Screening Mammography Abnormal Interpretation Rate (Recall Rate)* included in the American College of Radiology's National Radiology Data Registry (NRDR), were used to examine agreement between the measures' computed performance rates. To determine whether relationships exist among NQF #508, NQF #509, and ACRad5, the hypothesis was that good performance on one measure would be correlated to good performance on the other. Imaging on "probably benign" lesions result in unnecessary follow-up (NQF #508). Good performance indicated by not choosing an ambivalent response, e.g. "probably benign" assessment category, on NQF #508 likely indicates a physician who does not unnecessarily recall patients, in other words, has a low recall rate (ACRad5). Physicians who follow guidelines are likely in practices with good systems for tracking mammogram screening and remind patients of when follow-up mammograms are due (NQF #509). The following describes our analyses.

The computed performance data combined data from exams that occurred between 2014 and 2018, yielding 630 physicians who reported the three measures (NQF #508, NQF #509, and ACRad5).

Table 1.

Measure Identifier	Performance Rates				
	Mean	**Minimum**	**Median**	**Maximum**	**Standard Deviation**
ACRad5	10.03%	0%	9.06%	48.69%.	6.20%
NQF #508	1.14%	0%	0%	58.45%.	-
NQF #509	96.61%	0%	100%	100%	-

Table 2.

Measure Comparisons	Pearson Correlation Coefficients			
	Performance rates, unweighted	**Performance rates, weighted by number of exams**	**Performance rates, weighted by the square root of the exams total**	**Performance rates, weighted by the log of number of exams**
NQF #508 vs. NQF #509	0.026	0.0107	0.01923	0.02822
NQF #508 vs. ACRad5	-0.01876	-0.03356	-0.03642	-0.03395
NQF #509 vs ACRad5	-0.05152	-0.08987	-0.08584	-0.07078

Correlations were not detected between the performances rates of the measures.
All correlation coefficients are results of Pearson correlation.

Table 3 shows some variability in the measure by practice type and location.

Table 3.

	NQF #509		
	N	**Mean**	**Std Error**
	2572	99.02	0.16
Category			
Academic/university-based	271	100	0
Community hospital-based	7465	98.43	0.06

Freestanding imaging center	9439	99.79	0.04
Multi-specialty clinic	526	99.22	0.13
Other	3506	99.62	0.07
Location			
Metropolitan (> 100,000)	6889	99.65	0.06
Rural (<50,000)	3858	97.3	0.09
Suburban/Small (50,000-100,000)	10460	99.75	0.04
Region			
Midwest	3569	96.79	0.1
Northeast	2947	99.37	0.13
South	11834	99.84	0.03
West	2857	99.95	0.02

Feedback from a subset of experts confirmed that the face validity assessment described remains and that this measure retains strong face validity.

An expert panel was used to assess face validity of the measure. This panel consisted of 20 members, with representation from the ACR Commission on Breast Imaging and the National Mammography Database. The panel was asked to rate their agreement with the following statements:

1. The measure demonstrates a high impact on health care and an opportunity for improvement in quality over time.
2. Physicians who perform well on this measure demonstrate a higher level of quality than physicians who do not perform well on the measure.
3. In your opinion, how might this measure contribute to quality improvement? Check all that apply.
    - No value
    - Implementation of a reminder system could lead to an increase in mammography screening at appropriate intervals
    - Is complementary to the USPSTF guideline for screening mammograms to reduce breast cancer mortality
    - Promotes higher quality management and treatment

ACR Commission on Breast Imaging

Alson, Mark MD
Appleton, Catherine MD
Baker, Jay MD
Hendrick, R. Edward PhD
Lee, Carol MD
Monticciolo, Debra MD
Newell, Mary MD
Parkinson, Brett MD

ACR National Mammography Database Committee

Rosenberg, Robert MD
Sickles, Edward MD
Berg, Wendie MD
Ellis, Richard MD
Zuley, Margarita MD
Burnside, Elizabeth MD
Patel, Bhavika MD
Lee, Cindy

Rebner, Murray MD
Sickles, Edward MD
Smetherman, Dana MD
Smith, Robert PhD
Warren, Linda MD

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

The scores obtained from the measure as specified will accurately differentiate quality across providers.

Scale 1-5, where 1=Strongly Disagree; 3=Neither Disagree nor Agree; 5=Strongly Agree

The measure demonstrates a high impact on health care and an opportunity for improvement in quality over time.

Answer Options	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	Rating Average	Response Count
	0	0	0	3	7	4.70	10
						answered question	10
						skipped question	0

Physicians who perform well on this measure demonstrate a higher level of quality than physicians who do not perform well on the measure.

Answer Options	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	Rating Average	Response Count
	0	1	2	3	4	4.00	10
						answered question	10
						skipped question	0

In your opinion, how might this measure contribute to quality improvement? Check all that apply.

Answer Options	Response Percent	Response Count
No value	0.0%	0
Increase awareness of the appropriate use of mammographic assessment categories for screening mammography exams	90.0%	9

Is complementary to the recall rate metric used in Hospital Compare, with a 45-day period examined for recall	30%	3
Promotes higher quality management and treatment	70%	7
answered question		10
skipped question		0

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?*)

The expert panel agreed that the measure remained valid based on existing and new evidence.

1. **The measure demonstrates a high impact on health care and an opportunity for improvement in quality over time.**

   Responses to this statement were rated on a scale of 1 to 5, where 1 = Strongly Disagree and 5 = Strongly Agree. With 11 responses, the mean score was 4.09 which places the mean agreement between Agree and Strongly Agree. Only one respondent disagreed and no respondents strongly disagreed.

2. **Physicians who perform well on this measure demonstrate a higher level of quality than physicians who do not perform well on the measure.**

   Responses to this statement were rated on a scale of 1 to 5, where 1 = Strongly Disagree and 5 = Strongly Agree. With 11 responses, the mean score was 4.55 which places the mean agreement between Agree and Strongly Agree. No respondents were neutral and none disagreed or strongly disagreed.

3. **In your opinion, how might this measure contribute to quality improvement? Check all that apply.**

   Respondents to this question were able to choose any number of responses. Out of 11 respondents, 100% agreed that this measure would increase awareness of appropriate use, 54.5% believed it was complementary to the recall rate metric in Hospital Compare, and 81.8% believed it would promote higher quality management and treatment. No respondents felt that the measure had no value.

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)

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*)

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*)

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)*

To assess statistically significant differences in measure rates, the data described in sections above were used to calculate the mean, median, standard deviation, and interquartile range for the measure rates. In addition, the rates were divided into quartiles, and a Student's t-test was used to compare the rates of the plans in the 25th percentile to the rates of the plans in the 75th percentile.

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*)

Tables 4 and 5 below show the distribution of measure rates for **claims** data between 2015 and 2018. The mean rate was 92.21%, with a median rate of 100%, minimum rate of .05%, and maximum rate of 100%.

Table 4. Variation in Measure Rates for Claims Data – 2015 to 2018

Mean	Median	Standard Deviation
92.21%	100%	21.81%

Table 5. Distribution of Measure Rates for Claims Data – 2015 to 2018

Statistic	Value
Minimum	.05%
25th percentile	99.24%
50th percentile (median)	100 %
75th percentile	100%
Maximum	100%
Interquartile Range	.76%
Student's t-test p-value	P<.0001

Tables 6 and 7 below show the distribution of measure rates for **Registry** data between 2015 and 2018. The mean rate was 96.80%, with a median rate of 100%, minimum rate of .01%, and maximum rate of 100%.

Table 6. Variation in Measure Rates for Registry Data – 2015 to 2018

Mean	Median	Standard Deviation
96.80%	100%	14.10%

Table 7. Distribution of Measure Rates for Registry Data – 2015 to 2018

Statistic	Value
Minimum	.01%
25th percentile	100%

50th percentile (median)	100%
75th percentile	100%
Maximum	100%
Interquartile Range	0%
Student's t-test p-value	P<.0001

The tables below show the distribution of measure rates for **QCDR** data between 2015 and 2016. As a reminder, the QCDR data for 2017 and 2018 is combined in the registry data above. The mean rate was 99.99%, with a median rate of 100%, minimum rate of 99.28%, and maximum rate of 100%.

Table 8. Variation in Measure Rates for QCDR Data – 2015 to 2016

Mean	Median	Standard Deviation
99.99%	100%	.06%

Table 9. Distribution of Measure Rates for QCDR Data – 2015 to 2018

Statistic	Value
Minimum	99.28%
25th percentile	100%
50th percentile (median)	100%
75th percentile	100%
Maximum	100%
Interquartile Range	0%
Student's t-test p-value	P<.0001

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?*)
For the claims data, the measure rates did not show significant variation, with an interquartile range of .76%. There is no statistically significant difference in measure rates between the top and bottom quartile of the plans included in the testing (P< .0001 at alpha = 0.05). This variation shows that there is minor significant and clinically meaningful differences in rates across providers submitting claims information.

For the registry data, the measure rates did not show significant variation, with an interquartile range of 0%. There is no statistically significant difference in measure rates between the top and bottom quartile of the plans included in the testing (P<0.0001 at alpha = 0.05). This variation shows that there is minor statistically significant and clinically meaningful differences in rates across providers submitting registry information.

For the QCDR data, the measure rates did not show significant variation, with an interquartile range of 0%. There is no statistically significant difference in measure rates between the top and bottom quartile of the plans included in the testing (P<0.0001 at alpha = 0.05). This variation shows that there is minor statistically significant and clinically meaningful differences in rates across providers submitting QCDR information.

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*)

Not applicable.

With the use of claims and registry as the data sources for this measure, CMS Medicare and Medicaid administrative data is considered to largely be valid and reliable since it determines eligibility for enrollment and payment of services. Registry data may have some non-responders, as they are not required to submit all data to CMS. However, the volume of patients (68,844,412) used in this data set greatly minimizes the risk of bias.

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*)

Missing data related to registry data providers not submitting information on patients was previously noted. However, the amount of patients that were eligible (74,792,218) compared to the amount submitted and used for this analysis (68,844,412) likely would not have made a significant difference in the testing results.

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*)

Not applicable.

The performance results are from a significantly large data set of over 60,000,000 patients. The loss of ~5,000,000 eligible patients likely would not create a bias or a significant difference in the results. Each year, CMS raises the amount of data required for submission in the MIPS program. This will assist with minimizing bias even more in the future.

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)

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 health records (EHRs)

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).

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 was found to be reliable and feasible for implementation.

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)*.

Not applicable

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 Quality Improvement (Internal to the specific organization)	Payment Program Merit-based Incentives Payment System (MIPS) www.qpp.cms.gov MIPS www.qpp.cms.gov Merit-based Incentives Payment System (MIPS) www.qpp.cms.gov MIPS www.qpp.cms.gov

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

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

This measure has been included in the Physician Quality Reporting System since 2009 as Measure #225. Shown below are national average performance rates as reported in the CMS Report: 2013 Reporting Experience Including Trends (2007-2014) Physician Quality Reporting System and Electronic Prescribing (eRx) Incentive Program, APPENDIX, Table A27. Reporting and Performance Information by Individual Measure for the Physician Quality Reporting System (2010 to 2013).

Year Average Performance Rate
2010 N/A
2011 68.5 %
2012 74.6 %
2013 81.6%
2015 88.0%
2016 94.4%
2017 96.4%
2018 97.9%

The performance rate was calculated as the count of reported instances where performance was met (numerator) divided by the total number of reported instances that excluded reported exclusions (i.e., performance denominator).

The ACR believes that the reporting of participation information is a beneficial first step on a trajectory toward the public reporting of performance results, which is appropriate since the measure has been tested and the reliability of the performance data has been validated. Continued NQF endorsement will facilitate our ongoing progress toward this public reporting objective. Quality measures are tools that help measure health care processes and outcomes. These data are associated with the ability to provide high-quality health care and physician participation in quality programs such as MIPS.

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?*)

This is an accountability measure and used in the CMS quality and payment programs.

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.*)

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.

This is a MIPS measure currently in use through registries and claims. Detailed specifications are publicly available in the CMS Resource Library. Benchmarks are provided each year. MIPS reporters receive QRUR/MIPS reports. The current benchmark for the claims measure is 94.2-98.22 for decile 3, 98.23-99.67 for decile 4, 99.68-99.99 for decile 9 and 100 for decile 10. The registry measure has 100 for decile 10. Quarterly feedback reports are provided to QCDR users that report this measure. ACR staff is available to assist with the interpretation of this measure. This measure is mostly attributed to radiologists.

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.

ACR has been approved by CMS as a Qualified Clinical Data Registry since 2014. This measure is included in our portfolio of QCDR supported measures. Feedback is provided to all registry participants reporting any MIPS quality measure on a quarterly basis. Educational webinars are conducted monthly to explain measure requirements and interpretation of performance results.

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.

Overall, feedback is positive on this measure. Feedback is obtained from the registry and from CMS. The PQRS 2016 trend report shows Radiologists are in the top 10 specialty providers participating in PQRS via claims and registry. For claims 50.2% of our providers eligible for PQRS reporting participated in the program and 13.2% of providers participated through registries.

4a2.2.2. Summarize the feedback obtained from those being measured.

Feedback is obtained through our members, the CMS quality help desk, and CMS contractor QMMS.

4a2.2.3. Summarize the feedback obtained from other users

No other feedback has been provided on individuals not reporting the measure.

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.

This feedback is considered during the annual measure specification update process with CMS. The ACR Metrics Committee also review the feedback for annual updates.

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.

There is significant improvement from 2014 to 2018 for this measure.

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.

We are not aware of any unintended consequences related to this measurement.

4b2.2. Please explain any unexpected benefits from implementation of this measure.

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)

2372 : Breast Cancer Screening

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

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.

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.)

There are no competing measures (conceptually both the same measure focus and same target population).

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.

Available at measure-specific web page URL identified in S.1 **Attachment:**

Contact Information

Co.1 Measure Steward (Intellectual Property Owner): American College of Radiology

Co.2 Point of Contact: Judy, Burleson, jburleson@acr.org, 703-648-3787-

Co.3 Measure Developer if different from Measure Steward: American College of Radiology

Co.4 Point of Contact: Karen, Orozco, korozco@acr.org, 703-390-9848-

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.

List of Work Group Members:

William Golden, MD (Co-Chair) (internal medicine)

David Seidenwurm (Co-chair) (diagnostic radiology)

Michael Bettmann, MD

Dorothy Bulas, MD (pediatric radiology)

Rubin I. Cohen, MD, FACP, FCCP, FCCM

Richard T. Griffey, MD, MPH (emergency medicine)

Eric J. Hohenwalter, MD (vascular interventional radiology)

Deborah Levine, MD, FACR (radiology/ultrasound)

Mark Morasch, MD (vascular surgery)

Paul Nagy, MD, PhD (radiology)

Mark R. Needham, MD, MBA (family medicine)

Hoang D. Nguyen (diagnostic radiology/payer representative)

Charles J. Prestigiacomo, MD, FACS (neurosurgery)

William G. Preston, MD, FAAN (neurology)

Robert Pyatt, Jr., MD (diagnostic radiology)

Robert Rosenberg, MD (diagnostic radiology)

David A. Rubin, MD (diagnostic radiology)

B Winfred (B.W.) Ruffner, MD, FACP (medical oncology)

Frank Rybicki, MD, PhD, FAHA (diagnostic radiology)

Cheryl A. Sadow, MD (radiology)

John Schneider, MD, PhD (internal medicine)

Gary Schultz, DC, DACR (chiropractic)

Paul R. Sierzenski, MD, RDMS (emergency medicine)

Michael Wasylik, MD (orthopedic surgery)

Diagnostic Imaging Measure Development Work Group Staff

American College of Radiology: Judy Burleson, MHSA; Alicia Blakey, MS

American Medical Association-convened Physician Consortium for Performance Improvement: Mark Antman, DDS, MBA; Kathleen Blake, MD, MPH; Kendra Hanley, MS; Toni Kaye, MPH; Marjorie Rallins, DPM; Kimberly Smuk, RHIA; Samantha Tierney, MPH; Stavros Tsipas, MA

National Committee for Quality Assurance: Mary Barton, MD

PCPI measures are developed through cross-specialty, multi-disciplinary work groups. All medical specialties and other health care professional disciplines participating in patient care for the clinical condition or topic under study must be equal contributors to the measure development process. In addition, the PCPI strives to include on its work groups individuals representing the perspectives of patients, consumers, private health plans, and employers. This broad-based approach to measure development ensures buy-in on the measures from all stakeholders and minimizes bias toward any individual specialty or stakeholder group. All work groups have at least two co-chairs who have relevant clinical and/or measure development expertise and who are responsible for ensuring that consensus is achieved and that all perspectives are voiced.

Measure Developer/Steward Updates and Ongoing Maintenance

Ad.2 Year the measure was first released: 2007

Ad.3 Month and Year of most recent revision: 02, 2015

Ad.4 What is your frequency for review/update of this measure? These measures will be updated every 3 years.

Ad.5 When is the next scheduled review/update for this measure? 09, 2017

Ad.6 Copyright statement: The Measures are not clinical guidelines, do not establish a standard of medical care, and have not been tested for all potential applications.

The Measures, while copyrighted, can be reproduced and distributed, without modification, for noncommercial purposes, e.g., use by health care providers in connection with their practices. Commercial use is defined as the sale, license, or distribution of the Measures for commercial gain, or incorporation of the Measures into a product or service that is sold, licensed or distributed for commercial gain.

Commercial uses of the Measures require a license agreement between the user and the American Medical Association (AMA), [on behalf of the Physician Consortium for Performance Improvement® (PCPI®)] or American College of Radiology (ACR). Neither the AMA, ACR, PCPI, nor its members shall be responsible for any use of the Measures.

The AMA's, PCPI's and National Committee for Quality Assurance's significant past efforts and contributions to the development and updating of the Measures is acknowledged. ACR is solely responsible for the review and enhancement ("Maintenance") of the Measures as of December 31, 2014.

ACR encourages use of the Measures by other health care professionals, where appropriate.

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

Limited proprietary coding is contained in the Measure specifications for convenience. Users of the proprietary code sets should obtain all necessary licenses from the owners of these code sets. The AMA, ACR, the PCPI and its members disclaim all liability for use or accuracy of any Current Procedural Terminology (CPT®) or other coding contained in the specifications.

Ad.7 Disclaimers: See copyright statement above.

Ad.8 Additional Information/Comments: Coding/Specifications updates occur annually. The ACR has a formal measurement review process that stipulates regular (usually on a three-year cycle, when feasible) review of the measures. The process can also be activated if there is a major change in scientific evidence, results from testing or other issues are noted that materially affect the integrity of the measure.