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: Click to go to the link. ALT + LEFT ARROW to return

Purple text represents the responses from measure developers.

Red text denotes developer information that has changed since the last measure evaluation review.

Brief Measure Information

NQF #: 0034

Measure Title: Colorectal Cancer Screening (COL)

Measure Steward: National Committee for Quality Assurance

Brief Description of Measure: The percentage of patients 50–75 years of age who had appropriate screening for colorectal cancer.

Developer Rationale: This measure encourages screening for colorectal cancer so that it can be prevented or detected early when it is most treatable, which reduces deaths associated with colorectal cancer.

Numerator Statement: Patients who received one or more screenings for colorectal cancer according to clinical guidelines.

Denominator Statement: Patients 51–75 years of age

Denominator Exclusions: This measure excludes patients with a history of colorectal cancer or total colectomy. The measure also excludes patients who use hospice services or are enrolled in an institutional special needs plan (SNP) or living long-term in an institution any time during the measurement year.

Measure Type: Process

Data Source: Claims, Electronic Health Data, Paper Medical Records

Level of Analysis: Health Plan, Integrated Delivery System

Original Endorsement Date: Aug 10, 2009 **Most Recent Endorsement Date:** May 02, 2012

Staff 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**

Evidence Summary or **Summary of prior review in [year]**

- The 2017 United States Preventative Services Task Force (USPSTF) guidelines recommend screenings for colorectal cancer starting at age 50 and continuing until the age of 75. This guideline received an A rating, since the USPSTF concludes with high certainty that the benefits outweigh harms of performing colorectal cancer screening in patients age 50 to 75.
- The systematic review used to support this measure cites 47 articles (25 studies, fair or good quality) related to the effectiveness of screening programs based on the pre-specified screening tests (alone or in combination) in reducing incidence of and mortality from colorectal cancer; 44 articles (33 diagnostic accuracy studies, fair or good quality) related to the test performance characteristics of the pre-specified screening tests (alone or in combination) for detecting colorectal cancer, advanced adenomas, or adenomatous polyps based on size; and 113 articles (98 studies fair or good quality) related to the adverse effects of the different screening tests (either as single application or in a screening program) and variation in adverse effects by important subpopulations.
- The developer provides the following logic model for the measure: Adults at risk for colorectal cancer → Screening for colorectal cancer → Abnormal screening result → Evaluation and follow-up → Early detection and treatment of cancer → Improved length and/or quality of life

Changes to evidence from last review

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

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

Updates:

- The developer updated the Evidence form to provide the 2017 USPSTF guidelines (from the 2011 recommendation). Both the 2011 and 2017 guidelines recommend colorectal cancer screenings for individuals beginning at age 50 and continuing until age 75. Both guidelines received an A rating, meaning that the USPSTF recommends the service and there is high certainty that the net benefit is substantial.

Questions for the Committee:

- *The developer provided an updated 2017 USPSTF guideline to support the measure focus. Does the Committee agree that the measure reflects the current USPSTF recommendation?*

Guidance from the Evidence Algorithm

Measure a health outcome (Box 1) No → Assess performance of intermediate outcome, process, or structure(Box 3) Yes → Summary of QQC provided (Box 4) Yes → High certainty that the net benefit is substantial (Box 5) → High rating

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

1b. Gap in Care/Opportunity for Improvement and 1b. Disparities

Maintenance measures – increased emphasis on gap and variation

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

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

Commercial Plans (HMO and PPO combined)			
Measurement Year	**2014**	**2015**	**2016**
Mean	61.2%	60.0%	60.1%
Std. dev.	8.9%	9.2%	9.6%
10th percentile	50.4%	49.2%	48.4%
25th percentile	54.9%	54.1%	53.9%
50th percentile	60.3%	59.5%	60.1%
75th percentile	67.6%	66.3%	66.4%
90th percentile	72.0%	71.6%	72.2%
Interquartile range	12.7	12.2	12.5

Medicare Rates (HMO and PPO combined)			
Measurement year	**2014**	**2015**	**2016**
Mean	65.5%	67.2%	67.7%
Std. dev.	11.6%	10.9%	12.4%
10th percentile	51.6%	52.6%	50.8%
25th percentile	59.9%	60.9%	60.9%
50th percentile	66.9%	68.1%	69.9%
75th percentile	73.1%	74.5%	76.4%
90th percentile	77.4%	79.6%	81.0%
Interquartile range	13.2	13.7	15.5

- From 2014 to 2016, performance rates for this measure have been generally stable or shown some improvement.
- The developer provided the following data for the denominator for the performance data for 2014, 2015, and 2016.

Commercial			
Measurement year	2014	2015	2016
Number of plans	401	415	412
Median denominator size by plan	411	411	411

Medicare			
Measurement year	2014	2015	2016
Number of plans	401	415	412
Median denominator size by plan	411	411	411

Disparities

- HEDIS data are stratified by type of insurance (e.g. Commercial, Medicaid, Medicare). While not specified in the measure, this measure can also be stratified by demographic variables, such as race/ethnicity or socioeconomic status, in order to assess the presence of health care disparities, if the data are available to a plan.
- The developer provides disparities data from the literature, as follows:
  - Researchers have identified disparities in the rate of colorectal cancer screening based on race, ethnicity, income, education and English language proficiency. Racial/ethnic minorities, most notably Hispanic-Spanish, had lower colorectal cancer screening rates than Whites in 2010 (30.6% Hispanic-Spanish, 47.2% Asian, 49.5% American Indian/Alaska Native, 52.5% Hispanic-English, and 54.6% Native Hawaiian/Pacific Islander, compared to 62% White) (Liss and Baker, 2014).
  - Low-income populations have low colorectal cancer screening rates. The percentage of people who are up-to-date with screening has been consistently lower for people with a family income below 200 percent of the federal poverty level compared to people with a family income greater than or equal to 500 percent of the federal poverty level (In 2008, screening rate of 40.1% for people below 200 percent federal poverty level and 66.0% for people greater than or equal to 500 percent federal poverty level).
  - The percentage of people who are up-to-date with screening has been consistently lower for people with less than a high school education compared to people with greater than a high school education (screening rate of 37.5% in less than high school and 62.0% in greater than high school). (Klabunde et al, 2011)
  - Limited-English proficient populations exhibit lower colorectal cancer screening rates compared to English proficient populations. In 2006, 33% of Latinos responding in Spanish reported having been screened, compared to 51% of Latinos responding in English and 62% of English-speaking non-Latinos. (Diaz et al, 2008)

Citations

Diaz JA, Roberts MB, Goldman RE, Weitzen S, Eaton CB. Effect of language on colorectal cancer screening among latinos and non-latinos. Cancer epidemiology, biomarkers & prevention?: a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology. 2008;17(8)2169-2173.

Klabunde CN, Cronin KA, Breen N, Waldron WR, Ambs AH, Nadel MR. Trends in colorectal cancer test use among vulnerable populations in the U.S. Cancer epidemiology, biomarkers & prevention?: a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology. 2011;20(8):1611-1621.

Liss DT, Baker DW. Understanding current racial/ethnic disparities in colorectal cancer screening in the United States: the contribution of socioeconomic status and access to care. American Journal of Preventive Medicine. 2014;46(3):228-236.

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

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

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.

Complex measure evaluated by Scientific Methods Panel? ☐ **Yes** ☒ **No**

Evaluators: NQF Staff

Evaluation of Reliability and Validity (and composite construction, if applicable): Staff Scientific Acceptability Preliminary Analysis

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

Scientific Acceptability

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

Instructions:

- Please complete this form for each measure you are evaluating.
- Please pay close attention to the skip logic directions.
- If you are unable to check a box, please highlight or shade the box for your response.
- You must answer the "overall rating" item for both Reliability and Validity. Also, be sure to answer the composite measure question at the end of the form if your measure is a composite.
- We have provided TIPS to help you answer the questions.
- We've designed this form to try to minimize the amount of writing that you have to do. That said, **it is critical that you explain your thinking/rationale if you check boxes where we ask for an explanation** (because this is a Word document, you can just add your explanation below the checkbox). Feel free to add additional explanation, even if an explanation is not requested (but please type this underneath the appropriate checkbox).

- This form is based on Algorithms 2 and 3 in the Measure Evaluation Criteria and Guidance document (see pages 18-24). These algorithms provide guidance to help you rate the Reliability and Validity subcriteria. **We ask that you refer to this document when you are evaluating your measures.**
- Please contact Methods Panel staff if you have questions (methodspanel@qualityforum.org).

Measure Number: 0034

Measure Title: Colorectal Cancer Screening

RELIABILITY

1. Are submitted specifications precise, unambiguous, and complete so that they can be consistently implemented? *NOTE: NQF staff will conduct a separate, more technical, check of eMeasure (eCQM) specifications, value sets, logic, and feasibility, so no need to consider these in your evaluation.*

   TIPS: Consider the following: Are all the data elements clearly defined? Are all appropriate codes included? Is the logic or calculation algorithm clear? Is it likely this measure can be consistently implemented?

   ☒Yes (go to Question #2)

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

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

   TIPS: Check the 2nd "NO" box below if: only descriptive statistics provided; only describes process for data management/cleaning/computer programming; testing does not match measure specifications (i.e. data, eMeasure, level of analysis, patients)

   ☒Yes (go to Question #4)

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

3. Was **empirical VALIDITY testing** of patient-level data conducted?

   ☐Yes (use your rating from data element validity testing – Question #16- under Validity Section)

   ☐No (please explain below and rate Question #11: OVERALL RELIABILITY as INSUFFICIENT and proceed to the VALIDITY SECTION)

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

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

   ☒Yes (go to Question #5)

   ☐No (go to Question #8)

5. 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.*

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

   ☒Yes (go to Question #6)

   ☐No (please explain below then go to Question #8)

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

TIPS: Consider the following: Is the test sample adequate to generalize for widespread implementation? Do the results demonstrate sufficient reliability so that differences in performance can be identified?

☒High (go to Question #8)

☐Moderate (go to Question #8)

☐Low (please explain below then go to Question #7)

7. Was other reliability testing reported?

☐Yes (go to Question #8)

☐No (rate Question #11: OVERALL RELIABILITY as LOW and proceed to the VALIDITY SECTION)

8. Was reliability testing conducted with <u>patient-level data elements</u> that are used to construct the performance measure?

TIPS: Prior reliability studies of the same data elements may be submitted; if comparing abstraction to "authoritative source/gold standard" see Validity Section Question #15)

☐Yes (go to Question #9)

☒No (if there is score-level testing, rate Question #11: OVERALL RELIABILITY based on score-level rating from Question #6; otherwise, rate Question #11: OVERALL RELIABILITY as INSUFFICIENT. Then proceed to the VALIDITY SECTION)

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

TIPS: For example: inter-abstractor agreement (ICC, Kappa); other accepted method with description of how it assesses reliability of the data elements

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

☐Yes (go to Question #10)

☐No (if no, please explain below and rate Question #10 as INSUFFICIENT)

10. **RATING (data element)** – Based on the reliability statistic and scope of testing (number and representativeness of patients and entities), what is the level of certainty or confidence that the data used in the measure are reliable?

TIPS: Consider the following: Is the test sample adequate to generalize for widespread implementation? Can data elements be collected consistently?

☐Moderate (if score-level testing was NOT conducted, rate Question #11: OVERALL RELIABILITY as MODERATE)

☐Low (if score-level testing was NOT conducted, rate Question #11: OVERALL RELIABILITY as LOW)

☐Insufficient (go to Question #11)

11. OVERALL RELIABILITY RATING

OVERALL RATING OF RELIABILITY taking into account precision of specifications and <u>all</u> testing results:

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

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

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

☐Insufficient (please explain below) [NOTE: For most measure types, testing at both the score level and the data element level is <u>not</u> required]

VALIDITY

ASSESSMENT OF THREATS TO VALIDITY

1. Were all potential threats to validity that are relevant to the measure empirically assessed?

   TIPS: Threats to validity include: exclusions; need for risk adjustment; Able to identify statistically significant and meaningful differences; multiple sets of specifications; missing data/nonresponse.

   ☒Yes (go to Question #2)

   ☐No (please explain below and go to Question #2) [NOTE that even if ***non-assessment of applicable threats should result in an overall INSUFFICENT rating for validity***, we still want you to look at the testing results]

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

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

   ☐Yes (please explain below then go to Question #3)

   ☐No (go to Question #3)

   ☒Not applicable (i.e., there are no exclusions specified for the measure; go to Question #3)

3. Analysis of potential threats to validity: Risk-adjustment (applies to all outcome, cost, and resource use measures; may also apply to other types of measure)

   ☒Not applicable (e.g., structure or process measure that is not risk-adjusted; go to Question #4)

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

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

   c. Any concerns regarding the risk-adjustment approach?

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

   ☐Yes (please explain below then go to Question #4)

   ☐No (go to Question #4)

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

   ☐Yes (please explain below then go to Question #5)

   ☒No (go to Question #5)

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

   ☐Yes (please explain below then go to Question #6)

☐No (go to Question #6)

☒Not applicable (go to Question #6)

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

☐Yes (please explain below then go to Question #7)

☒No (go to Question #7)

ASSESSMENT OF MEASURE TESTING

7. Was empirical validity testing conducted using the measure as specified and appropriate statistical test?

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

☒Yes (go to Question #10) [NOTE: If appropriate empirical testing has been conducted, then evaluation of face validity is not necessary. Go to Question #8 **only if** there is insufficient information provided to evaluate data element and score-level testing.]

☐No (please explain below then go to Question #8)

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

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

☐Yes (go to Question #9)

☐No (please explain below and rate Question #17: OVERALL VALIDITY as INSUFFICIENT)

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

☐Yes (if a NEW measure, rate Question #17: OVERALL VALIDITY as MODERATE)

☐ Yes (if a MAINTENANCE measure, do you agree with the justification for not

conducting empirical testing? If no, rate Question #17: OVERALL VALIDITY as

INSUFFICIENT; otherwise, rate Question #17: OVERALL VALIDITY as MODERATE)

☐No (please explain below and rate Question #17: OVERALL VALIDITY AS LOW)

10. Was validity testing conducted with computed performance measure scores for each measured entity?

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

☒Yes (go to Question #11)

☐No (please explain below and go to Question #13)

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

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

☒Yes (go to Question #12)

☐No (please explain below, rate Question #12 as INSUFFICIENT and then go to Question #14)

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

    ☐High (go to Question #14)

    ☒Moderate (go to Question #14)

    ☐Low (please explain below then go to Question #13)

    ☐Insufficient

13. Was other validity testing reported?

    ☐Yes (go to Question #14)

    ☐No (please explain below and rate Question #17: OVERALL VALIDITY as LOW)

14. Was validity testing conducted with <u>patient-level data elements</u>?

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

    ☐Yes (go to Question #15)

    ☒No (please explain below and rate Question #17: OVERALL VALIDITY as INSUFFICIENT if <u>no</u> score-level testing was conducted, otherwise, rate Question #17: OVERALL VALIDITY based on score-level rating from Question #12)

15. 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.*

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

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

    ☐Yes (go to Question #16)

    ☐No (please explain below and rate Question #16 as INSUFFICIENT)

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

    ☐Moderate (if <u>score-level</u> testing was NOT conducted, rate Question #17: OVERALL VALIDITY as MODERATE)

    ☐Low (please explain below) (if <u>score-level</u> testing was NOT conducted, rate Question #17: OVERALL VALIDITY as LOW)

    ☐Insufficient (go to Question #17)

17. OVERALL VALIDITY RATING

OVERALL RATING OF VALIDITY taking into account the results and scope of <u>all</u> testing and analysis of potential threats.

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

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

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

☐Insufficient (if insufficient, please explain below) [NOTE: For most measure types, testing at both the score level and the data element level is not required] [NOTE: If rating is INSUFFICIENT for all empirical testing, then go back to Question #8 and evaluate any face validity that was conducted, then reconsider this overall rating.]

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

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 are generated or collected by and used by healthcare personnel during the provision of care (e.g., blood pressure, lab value, diagnosis, depression score), coded by someone other than person obtaining original information (e.g., DRG, ICD-9 codes on claims), abstracted from a record by someone other than person obtaining original information (e.g., chart abstraction for quality measure or registry)
- To allow for widespread reporting across health plans and health care practices, this measure is collected through multiple data sources (administrative data, electronic clinical data, paper records, and registry). The developer anticipates that as electronic health records become more widespread the reliance on paper record review will decrease.

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

Committee Pre-evaluation Comments: Criteria 3: Feasibility

Criterion 4: Usability and Use

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

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

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

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

Current uses of the measure

Publicly reported? ☒ **Yes** ☐ **No**

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

Accountability program details

- This measure is included in the composite Medicare Advantage Star Rating Program and is used in the Quality Payment Program (QPP).
- This measure is publicly reported nationally and by geographic regions in the NCQA State of Health Care annual report.

- This measure is used in the California P4P program, which is the largest non-governmental physician incentive program in the United States.
- This measure also is used in Quality Compass which is an tool used for selecting health plans, conducting competitor analysis, examining quality improvement and benchmarking plan performance, as well as the NCQA Health Plan Ratings/Report Card. The measure is used in NCQA accreditation for both Health Plans and Accountable Care Organizations (ACO) as well as the Qualified Health Plan (QHP) Quality Rating System.

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

Feedback on the measure by those being measured or others

- Questions received through the Policy Clarification Support system have generally centered around clarification on whether certain notations in medical record documentation are sufficient to meet measure criteria. Other questions have sought clarification about the screening methods that satisfy the measure numerator. During a recent public comment session, a majority of comments from measured entities supported updates to the measure to align with the latest clinical recommendations.
    - During the measure's last major update, feedback obtained through the NCQA feedback mechanisms resulted in specifications that include the new screening methods recommended by the USPSTF and other major clinical guideline organizations.

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

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

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

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

Improvement results

- From 2014 to 2016, performance rates for this measure have been generally stable or shown improvement. In 2016, commercial plans on average performance rate of 60%, and Medicare plans had an average rate of 68%.
- Given the updated USPSTF guidelines for colorectal cancer screening and the recent changes to this measure, the developer believes performance may improve in the coming years. In 2016, two additional screening methods were added to the guideline and measure. The developer hypothesizes that addition of more screening options may help patients feel more comfortable with the screening process, and therefore increase the number of patients who choose to be screened for colorectal cancer.

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 reported that no identified unintended consequences for this measure were identified during testing or since implementation.

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

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

Criterion 5: Related and Competing Measures

Related or competing measures

- 0658: Appropriate Follow-Up Interval for Normal Colonoscopy in Average Risk Patients (American Gastroenterological Association)
- Colorectal Cancer Screening – Minnesota Community Measurement (not NQF endorsed)

Harmonization

- The developer reports that the measure is harmonized to the extent possible.
- NQF #0658: Appropriate Follow-Up Interval for Normal Colonoscopy in Average Risk Patients focuses on only one of the available screening methods: colonoscopy. The measure assesses whether patients who have had a colonoscopy also have a recommended follow-up interval of 10 years documented in their colonoscopy report, whereas NQF #0034 focuses on several available screening methods in addition to colonoscopy.
- The Minnesota Community Measurement quality measure is intended for use at the clinician or practice-level, whereas NQF#0034 is intended for use at the health plan level.

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

Public and Member Comments

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

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

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

0034_-_Colorectal_Cancer_Screening__-_Evidence_7.1.docx

1a.1 For Maintenance of Endorsement: Is there new evidence about the measure since the last update/submission? Do not remove any existing information. If there have been any changes to evidence, the Committee will consider the new evidence. Please use the most current version of the evidence attachment (v7.1). Please use red font to indicate updated evidence.

Yes

1a Evidence (subcriterion 1a)

Measure Number (*if previously endorsed*)**:** 0034

Measure Title: Colorectal Cancer Screening

IF the measure is a component in a composite performance measure, provide the title of the Composite Measure here: N/A

Date of Submission: 11/15/2017

Instructions

- *Complete 1a.1 and 1a.2 for all measures. If instrument-based measure, complete 1a.3.*
- *Complete **EITHER 1a.2, 1a.3 or 1a.4** as applicable for the type of measure and evidence.*
- *For composite performance measures:*
  - *A separate evidence form is required for each component measure unless several components were studied together.*
  - *If a component measure is submitted as an individual performance measure, attach the evidence form to the individual measure submission.*
- All information needed to demonstrate meeting the evidence subcriterion (1a) must be in this form. An appendix of *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.

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

1a. Evidence to Support the Measure Focus

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

- Outcome: [3] Empirical data demonstrate a relationship between the outcome and at least one healthcare structure, process, intervention, or service. If not available, wide variation in performance can be used as evidence, assuming the data are from a robust number of providers and results are not subject to systematic bias.
- Intermediate clinical outcome: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured intermediate clinical outcome leads to a desired health outcome.

- Process: [5] a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured process leads to a desired health outcome.
- Structure: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured structure leads to a desired health outcome.
- Efficiency: [6] evidence not required for the resource use component.
- For measures derived from patient reports, evidence should demonstrate that the target population values the measured outcome, process, or structure and finds it meaningful.
- Process measures incorporating Appropriate Use Criteria: See NQF's guidance for evidence for measures, in general; guidance for measures specifically based on clinical practice guidelines apply as well.

Notes

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

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

5. Clinical care processes typically include multiple steps: assess → identify problem/potential problem → choose/plan intervention (with patient input) → provide intervention → evaluate impact on health status. If the measure focus is one step in such a multistep process, the step with the strongest evidence for the link to the desired outcome should be selected as the focus of measurement. Note: A measure focused only on collecting PROM data is not a PRO-PM.

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

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

☐ Outcome:

☐Patient-reported outcome (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*):

☒ Process: Colorectal cancer screening

☐ Appropriate use measure:

☐ Structure:

☐ Composite:

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.

Adults at risk for colorectal cancer >>> Screening for colorectal cancer >>> Abnormal screening result >>> Evaluation and follow-up >>> Early detection and treatment of cancer >>> Improved length and/or quality of life

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

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

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

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

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

☐ Clinical Practice Guideline recommendation (with evidence review)

☒ US Preventive Services Task Force Recommendation

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

☐ Other

Source of Systematic Review: • Title • Author • Date • Citation, including page number • URL	2017 Submission **US Preventive Services Task Force (USPSTF). 2016. "Screening for Colorectal Cancer: US Preventive Services Task Force Recommendation Statement." *JAMA* 315(23):2564-2575. doi: 10.1001/jama.2016.5989** 2011 Submission **US Preventive Services Task Force (USPSTF). Screening for colorectal cancer: U.S. Preventive Services Task Force recommendation statement. Ann Intern Med 2008 Nov 4;149(9):627-37. http://www.uspreventiveservicestaskforce.org/uspstf/uspscolo.htm**
Quote the guideline or recommendation verbatim about the process, structure or intermediate outcome being measured. If not a guideline, summarize the conclusions from the SR.	**2017 Submission** "The USPSTF recommends screening for colorectal cancer starting at age 50 years and continuing until age 75 years (A recommendation)" **2011 Submission** The USPSTF recommends screening for colorectal cancer in adults, beginning at age 50 years and continuing until age 75 years. (A recommendation)
Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	**2017 Submission** The USPSTF concludes with high certainty that the benefits outweigh harms of performing colorectal cancer screening in patients age 50 to 75.
Provide all other grades and definitions from the evidence grading system	**2017 Submission** N/A
Grade assigned to the **recommendation** with definition of the grade	**2017 Submission** Grade: A "The USPSTF recommends the service. There is high certainty that the net benefit is substantial." **2011 Submission** Grade: A The USPSTF recommends the service. There is high certainty that the net benefit is substantial.

Provide all other grades and definitions from the recommendation grading system	**2017 Submission** 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. 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. 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. 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. **2011 Submission** 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. C. The USPSTF recommends against routinely providing the service. There may be considerations that support providing the service in an individual patient. There is at least moderate certainty that the net benefit is small. 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. 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.

Body of evidence: • Quantity – how many studies? • Quality – what type of studies?	**2017 Submission** The evidence report supporting this guideline outlines the quantity and quality of evidence (Lin et al 2016). Key question 1: What is the effectiveness of screening programs based on the pre-specified screening tests (alone or in combination) in reducing incidence of and mortality from colorectal cancer? • Included 47 articles (25 studies, fair or good quality) Key question 2: What are the test performance characteristics of the prespecified screening tests (alone or in combination) for detecting colorectal cancer, advanced adenomas, or adenomatous polyps based on size? • Included 44 articles (33 diagnostic accuracy studies, fair or good quality) Key question 3a: What are the adverse effects of the different screening tests (either as single application or in a screening program)? Key Question 3b: Do adverse effects vary by important subpopulations (eg, age)? • Included 113 articles (98 studies, fair or good quality) Lin, J.S., M.A. Piper, L.A. Perdue, et al. 2016. "Screening for Colorectal Cancer: Updated Evidence Report and Systematic Review for the US Preventive Services Task Force." *JAMA* 315(23):2576-94. doi: 10.1001/jama.2016.3332. **2011 Submission** Quantity: Refer to USPSTF http://www.uspreventiveservicestaskforce.org/uspstf08/colocancer/coloartwhit.htm Quality: High
Estimates of benefit and consistency across studies	**2017 Submission** The USPSTF recommendation states: "The USPSTF concludes with high certainty that screening for colorectal cancer in average-risk, asymptomatic adults aged 50 to 75 years is of substantial net benefit. Multiple screening strategies are available to choose from, with different levels of evidence to support their effectiveness, as well as unique advantages and limitations, although there are no empirical data to demonstrate that any of the reviewed strategies provide a greater net benefit. Screening for colorectal cancer is a substantially underused preventive health strategy in the United States." **2011 Submission** The available evidence usually includes consistent results from well-designed, well-conducted studies in representative primary care populations. These studies assess the effects of the preventive service on health outcomes. This conclusion is therefore unlikely to be strongly affected by the results of future studies.

What harms were identified?	**2017 Submission** The USPSTF guideline (2016) summarizes the harms of screening and early intervention: "The harms of screening for colorectal cancer in adults aged 50 to 75 years are small. The majority of harms result from the use of colonoscopy, either as the screening test or as follow-up for positive findings detected by other screening tests. The rate of serious adverse events from colorectal cancer screening increases with age."
Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	**2017 Submission** To our knowledge, there have been no published studies since the systematic review that would impact the recommendations.

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.

This measure encourages screening for colorectal cancer so that it can be prevented or detected early when it is most treatable, which reduces deaths associated with colorectal cancer.

1b.2. Provide performance scores on the measure as specified (current and over time) at the specified level of analysis. *(This is required for maintenance of endorsement. Include mean, std dev, min, max, interquartile range, scores by decile. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities include.) This information also will be used to address the sub-criterion on improvement (4b1) under Usability and Use.*

The following data are extracted from HEDIS data collection reflecting the most recent years of measurement for this measure. Performance data are summarized at the health plan level and summarized by mean, standard deviation, minimum health plan performance, maximum health plan performance and performance at 10th, 25th, 50th, 75th, and 90th percentile. Data are stratified by year and product line (i.e. commercial, Medicare).

Colorectal Cancer Screening – commercial Rate (HMO and PPO Combined)

MEASUREMENT YEAR	MEAN	ST DEV	10TH	25TH	50TH	75TH	90TH	Interquartile Range
2014	61.2%	8.9%	50.4%	54.9%	60.3%	67.6%	72.0%	12.7

2015	60.0%	9.2%	49.2%	54.1%	59.5%	66.3%	71.6%	12.2
2016	60.1%	9.6%	48.4%	53.9%	60.1%	66.4%	72.2%	12.5

Colorectal Cancer Screening – Medicare Rate (HMO and PPO Combined)

MEASUREMENT YEAR	MEAN	ST DEV	10TH	25TH	50TH	75TH	90TH	Interquartile Range
2014	65.5%	11.6%	51.6%	59.9%	66.9%	73.1%	77.4%	13.2
2015	67.2%	10.9%	52.6%	60.9%	68.1%	74.5%	79.6%	13.7
2016	67.7%	12.4%	50.8%	60.9%	69.9%	76.4%	81.0%	15.5

The data references are extracted from HEDIS data collection reflecting the most recent years of measurement for this measure. In 2016, HEDIS measures covered 114.2 million commercial health plan beneficiaries and 17.6 million Medicare beneficiaries. Below is a description of the denominator for this measure. It includes the number of health plans included in HEDIS data collection and the mean eligible population for the measure across health plans.

Colorectal Cancer Screening – commercial

YEAR	N Plans	Median Denominator Size per plan
2014	401	411
2015	415	411
2016	412	411

Colorectal Cancer Screening – Medicare

YEAR	N Plans	Median Denominator Size per plan
2014	449	396
2015	440	408
2016	459	411

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.

1b.4. Provide disparities data from the measure as specified (current and over time) by population group, e.g., by race/ethnicity, gender, age, insurance status, socioeconomic status, and/or disability. *(This is required for maintenance of endorsement. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included.) For measures that show high levels of performance, i.e., "topped out", disparities data may demonstrate an opportunity for improvement/gap in care for certain sub-populations. This information also will be used to address the sub-criterion on improvement (4b1) under Usability and Use.*

HEDIS data are stratified by type of insurance (e.g. Commercial, Medicaid, Medicare). While not specified in the measure, this measure can also be stratified by demographic variables, such as race/ethnicity or socioeconomic status, in order to assess the presence of health care disparities, if the data are available to a plan. The HEDIS Race/Ethnicity Diversity of Membership and the Language Diversity of Membership measures were designed to promote standardized methods for collecting these data and follow Office of Management and Budget and Institute of Medicine guidelines for collecting and categorizing race/ethnicity and language data. In addition, NCQA's Multicultural Health Care Distinction Program outlines standards for collecting, storing, and using race/ethnicity and language data to assess health care disparities. Based on extensive work by NCQA to understand how to promote culturally and linguistically appropriate services among plans and providers, we have many examples of how health plans have used HEDIS measures to design quality improvement programs to decrease disparities in care.

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

Researchers have identified disparities in the rate of colorectal cancer screening based on race, ethnicity, income, education and English language proficiency. Racial/ethnic minorities, most notably Hispanic-Spanish, had lower colorectal cancer screening rates than Whites in 2010 (30.6 percent Hispanic-Spanish, 47.2 percent Asian, 49.5 percent American Indian/Alaska Native, 52.5 percent Hispanic-English and 54.6 percent Native Hawaiian/Pacific Islander, compared to 62 percent White) (Liss and Baker, 2014). Low-income and low-literacy populations also have low colorectal cancer screening rates. The percentage of people who are up-to-date with screening has been consistently lower for people with a family income below 200 percent of the federal poverty level compared to people with a family income greater than or equal to 500 percent of the federal poverty level (In 2008, screening rate of 40.1 percent for people below 200 percent federal poverty level and 66.0 percent for people greater than or equal to 500 percent federal poverty level). Similarly, the percentage of people who are up-to-date with screening has been consistently lower for people with less than a high school education compared to people with greater than a high school education (screening rate of 37.5 percent in less than high school and 62.0 percent in greater than high school). (Klabunde et al, 2011) Limited-English proficient populations exhibit lower colorectal cancer screening rates compared to English proficient populations. In 2006, 33 percent of Latinos responding in Spanish reported having a screen, compared to 51 percent of Latinos responding in English and 62 percent of English-speaking non-Latinos. (Diaz et al, 2008)

Brenner AT, Hoffman R, McWilliams A, Pignone MP, Rhyne RL, Tapp H, Weaver MA, Callan D, de Hernandez BU, Harbi K, Reuland DS. Colorectal cancer screening in vulnerable patients: promoting informed and shared decisions. American Journal of Preventive Medicine. 2016;51(4)454-462.

Diaz JA, Roberts MB, Goldman RE, Weitzen S, Eaton CB. Effect of language on colorectal cancer screening among latinos and non-latinos. Cancer epidemiology, biomarkers & prevention?: a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology. 2008;17(8)2169-2173.

Klabunde CN, Cronin KA, Breen N, Waldron WR, Ambs AH, Nadel MR. Trends in colorectal cancer test use among vulnerable populations in the U.S. Cancer epidemiology, biomarkers & prevention?: a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology. 2011;20(8):1611-1621.

Liss DT, Baker DW. Understanding current racial/ethnic disparities in colorectal cancer screening in the United States: the contribution of socioeconomic status and access to care. American Journal of Preventive Medicine. 2014;46(3):228-236.

Rice K, Gressard L, DeGroff A, Gersten J, Robie, J, Leadbetter S, Glover-Kudon R, Butterly L. Increasing colonoscopy screening in disparate populations: results from an evaluation of patient navigation in the New Hampshire Colorectal Cancer Screening Program. Cancer. 2017;123(17)3356-3366.

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 : Colorectal

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

Primary Prevention

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

Elderly, Populations at Risk : Dual eligible beneficiaries

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

N/A

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

This is not an eMeasure **Attachment:**

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

Attachment **Attachment:** 0034_COL_Value_Sets.xlsx

S.2c. Is this an instrument-based measure (i.e., data collected via instruments, surveys, tools, questionnaires, scales, etc.)? Attach copy of instrument if available.

No, this is not an instrument-based measure **Attachment:**

S.2d. Is this an instrument-based measure (i.e., data collected via instruments, surveys, tools, questionnaires, scales, etc.)? Attach copy of instrument if available.

Not an instrument-based measure

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

Yes

S.3.2. <u>For maintenance of endorsement,</u> please briefly describe any important changes to the measure specifications since last measure update and explain the reasons.

Since the last NQF review, two additional screening methods have been added to the measure, in alignment with updates to clinical guidelines. These changes were reviewed by stakeholder groups, vetted through a public comment period, and approved by our committees.

S.4. Numerator Statement *(Brief, narrative description of the measure focus or what is being measured about the target population, i.e., cases from the target population with the target process, condition, event, or outcome) DO NOT include the rationale for the measure.*

<u>IF an OUTCOME MEASURE</u>, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm (S.14).

Patients who received one or more screenings for colorectal cancer according to clinical guidelines.

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)*

<u>IF an OUTCOME MEASURE,</u> describe how the observed outcome is identified/counted. Calculation of the risk-adjusted outcome should be described in the calculation algorithm (S.14).

ADMINISTRATIVE:

Patients who received one or more screenings for colorectal cancer. Any of the following meet criteria:

-Fecal occult blood test (FOBT Value Set) during the measurement year.

-Flexible sigmoidoscopy (Flexible Sigmoidoscopy Value Set) during the measurement year or the four years prior to the measurement year.

-Colonoscopy (Colonoscopy Value Set) during the measurement year or the nine years prior to the measurement year.

-CT colonography (CT Colonography Value Set) during the measurement year or the four years prior to the measurement year.

-FIT-DNA test (FIT-DNA Value Set) during the measurement year or the two years prior to the measurement year.

MEDICAL RECORD:

Patients who received one or more screenings for colorectal cancer. Any of the following meet criteria:

-Fecal occult blood test during the measurement year.

-Flexible sigmoidoscopy during the measurement year or the four years prior to the measurement year.

-Colonoscopy during the measurement year or the nine years prior to the measurement year.

-CT colonography during the measurement year or the four years prior to the measurement year.

-FIT-DNA test during the measurement year or the two years prior to the measurement year.

Documentation in the medical record must include a note indicating the date when the colorectal cancer screening was performed. A result is not required if the documentation is clearly part of the “medical history” section of the record; if this is not clear, the result or finding must also be present (this ensures that the screening was performed and not merely ordered).

A pathology report that indicates the type of screening (e.g., colonoscopy, flexible sigmoidoscopy) and the date when the screening was performed meets criteria.

For pathology reports that do not indicate the type of screening and for incomplete procedures:

--Evidence that the scope advanced beyond the splenic flexure meets criteria for a completed colonoscopy.

--Evidence that the scope advanced into the sigmoid colon meets criteria for a completed flexible sigmoidoscopy.

There are two types of FOBT tests: guaiac (gFOBT) and immunochemical (FIT). Depending on the type of FOBT test, a certain number of samples are required for numerator compliance. Follow the instructions below to determine member compliance.

--If the medical record does not indicate the type of test and there is no indication of how many samples were returned, assume the required number was returned. The member meets the screening criteria for inclusion in the numerator.

--If the medical record does not indicate the type of test and the number of returned samples is specified, the member meets the screening criteria only if the number of samples specified is greater than or equal to three samples. If there are fewer than three samples, the member does not meet the screening criteria for inclusion.

--FIT tests may require fewer than three samples. If the medical record indicates that an FIT was done, the member meets the screening criteria, regardless of how many samples were returned.

--If the medical record indicates that a gFOBT was done, follow the scenarios below.

–If the medical record does not indicate the number of returned samples, assume the required number was returned. The member meets the screening criteria for inclusion in the numerator.

–If the medical record indicates that three or more samples were returned, the member meets the screening criteria for inclusion in the numerator.

–If the medical record indicates that fewer than three samples were returned, the member does not meet the screening criteria.

Do not count digital rectal exams (DRE), FOBT tests performed in an office setting or performed on a sample collected via DRE.

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

Patients 51–75 years of age

S.7. Denominator Details *(All information required to identify and calculate the target population/denominator such as definitions, time period for data collection, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b.)*

<u>IF an OUTCOME MEASURE</u>, describe how the target population is identified. Calculation of the risk-adjusted outcome should be described in the calculation algorithm (S.14).

Patients 51–75 years of age as of the end of the measurement year (e.g. December 31).

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

This measure excludes patients with a history of colorectal cancer or total colectomy. The measure also excludes patients who use hospice services or are enrolled in an institutional special needs plan (SNP) or living long-term in an institution any time during the measurement year.

S.9. Denominator Exclusion Details *(All information required to identify and calculate exclusions from the denominator such as definitions, time period for data collection, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b.)*

Exclude patients with either of the following any time during the patient's history through December 31 of the measurement year:

- Colorectal cancer (Colorectal Cancer Value Set)

- Total colectomy (Total Colectomy Value Set)

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

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

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

None

S.11. Risk Adjustment Type (Select type. Provide specifications for risk stratification in measure testing attachment)

No risk adjustment or risk stratification

If other:

S.12. Type of score:

Rate/proportion

If other:

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

Better quality = Higher score

S.14. Calculation Algorithm/Measure Logic (*Diagram or describe the calculation of the measure score as an ordered sequence of steps including identifying the target population; exclusions; cases meeting the target process, condition, event, or outcome; time period for data, aggregating data; risk adjustment; etc.)*

Step 1. Determine the eligible population: identify patients 51-75 years of age by the end of the measurement year.

Step 2. Search for an exclusion in the patient's history: history of total colectomy or colorectal cancer. Exclude these patients from the eligible population.

Step 3. Determine numerator: the number of patients who have been screened for colorectal cancer by any of the included screening methods, within the associated time interval.

Step 4. Calculate the rate.

S.15. Sampling *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*

IF an instrument-based performance measure (e.g., PRO-PM), identify whether (and how) proxy responses are allowed.

N/A

S.16. Survey/Patient-reported data *(If measure is based on a survey or instrument, provide instructions for data collection and guidance on minimum response rate.)*

Specify calculation of response rates to be reported with performance measure results.

N/A

S.17. Data Source *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*

If other, please describe in S.18.

Claims, Electronic Health Data, Paper Medical Records

S.18. Data Source or Collection Instrument *(Identify the specific data source/data collection instrument (e.g. name of database, clinical registry, collection instrument, etc., and describe how data are collected.)*

IF instrument-based, identify the specific instrument(s) and standard methods, modes, and languages of administration.

This measure is based on administrative claims and medical record documentation collected in the course of providing care to health plan members. NCQA collects the Healthcare Effectiveness Data and Information Set (HEDIS) data for this measure directly from Health Management Organizations and Preferred Provider Organizations via NCQA's online data submission system.

S.19. Data Source or Collection Instrument *(available at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*

No data collection instrument provided

S.20. Level of Analysis *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*

Health Plan, Integrated Delivery System

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

Outpatient Services

If other:

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

N/A

2. Validity – See attached Measure Testing Submission Form

0034_-_Colorectal_Cancer_Screening__-_Testing_7.1-636463498807302646.docx

2.1 For maintenance of endorsement

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

Yes

2.2 For maintenance of endorsement

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

Yes

2.3 For maintenance of endorsement

Risk adjustment: For outcome, resource use, cost, and some process measures, risk-adjustment that includes social risk factors is not prohibited at present. Please update sections 1.8, 2a2, 2b1,2b4.3 and 2b5 in the Testing attachment and S.140 and S.11 in the online submission form. NOTE: These sections must be updated even if social risk factors are not included in the risk-adjustment strategy. You MUST use the most current version of the Testing Attachment (v7.1) -- older versions of the form will not have all required questions.

No - This measure is not risk-adjusted

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

Measure Number (*if previously endorsed*)**:** 0034
Measure Title: Colorectal Cancer Screening
Date of Submission: 11/15/2017

Type of Measure:

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

Instructions

- Measures must be tested for all the data sources and levels of analyses that are specified. ***If there is more than one set of data specifications or more than one level of analysis, contact NQF staff*** about how to present all the testing information in one form.
- **For all measures, sections 1, 2a2, 2b1, 2b2, and 2b4 must be completed.**
- **For outcome and resource use measures**, section **2b3** also must be completed.
- If specified for **multiple data sources/sets of specificaitons** (e.g., claims and EHRs), section **2b5** also must be completed.
- Respond to all questions as instructed with answers immediately following the question. All information on testing to demonstrate meeting the subcriteria for reliability (2a2) and validity (2b1-2b6) must be in this form. An appendix for *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 25 pages (*incuding questions/instructions;* minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.
- For information on the most updated guidance on how to address social risk factors variables and testing in this form refer to the release notes for version 7.1 of the Measure Testing Attachment.

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

2a2. Reliability testing [10] demonstrates the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise. For **instrument-based measures** (including PRO-PMs) **and composite performance measures**, reliability should be demonstrated for the computed performance score.

2b1. Validity testing [11] demonstrates that the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality. For **instrument-based measures**

(including PRO-PMs) and composite performance measures, validity should be demonstrated for the computed performance score.

2b2. Exclusions are supported by the clinical evidence and are of sufficient frequency to warrant inclusion in the specifications of the measure; [12]

AND

If patient preference (e.g., informed decisionmaking) is a basis for exclusion, there must be evidence that the exclusion impacts performance on the measure; in such cases, the measure must be specified so that the information about patient preference and the effect on the measure is transparent (e.g., numerator category computed separately, denominator exclusion category computed separately). [13]

2b3. For outcome measures and other measures when indicated (e.g., resource use):

- **an evidence-based risk-adjustment strategy** (e.g., risk models, risk stratification) is specified; is based on patient factors (including clinical and social risk factors) that influence the measured outcome and are present at start of care; [14,15] and has demonstrated adequate discrimination and calibration

OR

- rationale/data support no risk adjustment/ stratification.

2b4. Data analysis of computed measure scores demonstrates that methods for scoring and analysis of the specified measure allow for **identification of statistically significant and practically/clinically meaningful** [16] **differences in performance**;

OR

there is evidence of overall less-than-optimal performance.

2b5. If multiple data sources/methods are specified, there is demonstration they produce comparable results.

2b6. Analyses identify the extent and distribution of **missing data** (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias.

Notes

10. Reliability testing applies to both the data elements and computed measure score. Examples of reliability testing for data elements include, but are not limited to: inter-rater/abstractor or intra-rater/abstractor studies; internal consistency for multi-item scales; test-retest for survey items. Reliability testing of the measure score addresses precision of measurement (e.g., signal-to-noise).

11. Validity testing applies to both the data elements and computed measure score. Validity testing of data elements typically analyzes agreement with another authoritative source of the same information. Examples of validity testing of the measure score include, but are not limited to: testing hypotheses that the measures scores indicate quality of care, e.g., measure scores are different for groups known to have differences in quality assessed by another valid quality measure or method; correlation of measure scores with another valid indicator of quality for the specific topic; or relationship to conceptually related measures (e.g., scores on process measures to scores on outcome measures). Face validity of the measure score as a quality indicator may be adequate if accomplished through a systematic and transparent process, by identified experts, and explicitly addresses whether performance scores resulting from the measure as specified can be used to distinguish good from poor quality. The degree of consensus and any areas of disagreement must be provided/discussed.

12. Examples of evidence that an exclusion distorts measure results include, but are not limited to: frequency of occurrence, variability of exclusions across providers, and sensitivity analyses with and without the exclusion.

13. Patient preference is not a clinical exception to eligibility and can be influenced by provider interventions.

14. Risk factors that influence outcomes should not be specified as exclusions.

15. With large enough sample sizes, small differences that are statistically significant may or may not be practically or clinically meaningful. The substantive question may be, for example, whether a statistically significant difference of one percentage point in the percentage of patients who received smoking cessation counseling (e.g., 74 percent v. 75 percent) is clinically meaningful; or whether a statistically significant difference of $25 in cost for an episode of care (e.g., $5,000 v. $5,025) is practically meaningful. Measures with overall less-than-optimal performance may not demonstrate much variability across providers.

1. DATA/SAMPLE USED FOR ALL TESTING OF THIS MEASURE

Often the same data are used for all aspects of measure testing. In an effort to eliminate duplication, the first five questions apply to all measure testing. If there are differences by aspect of testing,(e.g., reliability vs. validity) be sure to indicate the specific differences in question 1.7.

1.1. What type of data was used for testing? (*Check all the sources of data identified in the measure specifications and data used for testing the measure. Testing must be provided for all the sources of data specified and intended for measure implementation.* ***If different data sources are used for the numerator and denominator, indicate N [numerator] or D [denominator] after the checkbox.***)

Measure Specified to Use Data From: (*must be consistent with data sources entered in S.17*)	Measure Tested with Data From:
☒ abstracted from paper record	☒ abstracted from paper record
☒ claims	☒ claims
☐ registry	☐ registry
☐ abstracted from electronic health record	☐ abstracted from electronic health record
☐ eMeasure (HQMF) implemented in EHRs	☐ eMeasure (HQMF) implemented in EHRs
☐ other:	☐ other:

1.2. If an existing dataset was used, identify the specific dataset (*the dataset used for testing must be consistent with the measure specifications for target population and healthcare entities being measured; e.g., Medicare Part A claims, Medicaid claims, other commercial insurance, nursing home MDS, home health OASIS, clinical registry*).

N/A

1.3. What are the dates of the data used in testing? 2017 Submission: 2016 2011 Submission: 2009

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:	☐ other:

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*)

2017 Submission

Sample for measure score reliability testing: The measure score reliability was calculated from HEDIS data that included 459 Medicare health plans and 412 commercial health plans. The sample included all Medicare and commercial health plans submitting data to NCQA for HEDIS. The plans were geographically diverse and varied in size.

Sample for Construct Validity Testing: Construct validity was calculated from HEDIS data that included 430 Medicare health plans and 412 commercial health plans. The sample included all Medicare and commercial health plans submitting data to NCQA for HEDIS. The plans were geographically diverse and varied in size.

2011 Submission

HEDIS Health Plan performance data 2010

1.6. How many and which patients were included in the testing and analysis (by level of analysis and data source)? (*identify the number and descriptive characteristics of patients included in the analysis (e.g., age, sex, race, diagnosis); if a sample was used, describe how patients were selected for inclusion in the sample*)

2017 Submission

Patient sample for measure score reliability testing: In 2016, HEDIS measures covered 114.2 million commercial health plan beneficiaries and 17.6 million Medicare beneficiaries. Data are summarized at the health plan level and stratified by product line (i.e. commercial, Medicare). Below is a description of the sample. It includes number of health plans included HEDIS data collection and the median eligible population for the measure across health plans.

Product Type	Number of Plans	Median number of eligible patients per plan
Commercial	412	411
Medicare	459	411

Beneficiary Sample for Construct Validity Testing: In 2016, HEDIS measures covered 114.2 million commercial health plan beneficiaries and 17.6 million Medicare beneficiaries. Data is summarized at the health plan level. Data are stratified by product line (i.e. commercial, Medicare). Below is a description of the sample. It includes number of health plans included HEDIS data collection and the median eligible population for the measure across health plans.

Product Type	Number of plans	Median number of eligible patients per plan
Commercial	412	411
Medicare	430	411

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.

Reliability of the measure score was tested using a beta-binomial calculation. This analysis included the entire HEDIS data sample (described above).

Validity was demonstrated through construct validity.

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.

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

____

2a2. RELIABILITY TESTING

***Note**: If accuracy/correctness (validity) of data elements was empirically tested, separate reliability testing of data elements is not required – in 2a2.1 check critical data elements; in 2a2.2 enter "see section 2b2 for validity testing of data elements"; and skip 2a2.3 and 2a2.4.*

2a2.1. What level of reliability testing was conducted? (*may be one or both levels*)

☐ **Critical data elements used in the measure** (*e.g., inter-abstractor reliability; data element reliability must address ALL critical data elements*)

☒ **Performance measure score** (e.g., *signal-to-noise analysis*)

2a2.2. For each level checked above, describe the method of reliability testing and what it tests (*describe the steps—do not just name a method; what type of error does it test; what statistical analysis was used*)

2017 Submission

Reliability Testing of Performance Measure Score: same as below

2011 Submission

Reliability was estimated by using the beta-binomial model. Beta-binomial is a better fit when estimating the reliability of simple pass/fail rate measures as is the case with most HEDIS® health plan measures. The beta-binomial model assumes the plan score is a binomial random variable conditional on the plan´s true value that comes from the beta distribution. The beta distribution is usually defined by two parameters, alpha and beta. Alpha and beta can be thought of as intermediate calculations to get to the needed variance estimates. The beta distribution can be symmetric, skewed or even U-shaped.

Reliability used here is the ratio of signal to noise. The signal in this case is the proportion of the variability in measured performance that can be explained by real differences in performance. A reliability of zero implies that all the variability in a measure is attributable to measurement error. A reliability of one implies that all the variability is attributable to real differences in performance. The higher the reliability score, the greater is the confidence with which one can distinguish the performance of one plan from another. A reliability score greater than or equal to 0.7 is considered very good.

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*)

2017 Submission

Beta-Binomial Statistic:

Commercial	Medicare
0.997	0.988

2011 Submission

Commercial Plans 2010: reliability 0.994468

Medicaid 2010: Not available

Medicare 2010: reliability 0.993543

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?*)

2017 Submission

Interpretation of measure score reliability testing: The testing suggests the measure has high reliability.

____

2b1. VALIDITY TESTING

2b1.1. What level of validity testing was conducted? (*may be one or both levels*)

☐ **Critical data elements** (*data element validity must address ALL critical data elements*)

☒ **Performance measure score**

☒ **Empirical validity testing**

☒ **Systematic assessment of face validity of performance measure score as an indicator** of quality or resource use (*i.e., is an accurate reflection of performance on quality or resource use and can distinguish good from poor performance*) **NOTE**: Empirical validity testing is expected at time of maintenance review; if not possible, justification is required.

2b1.2. For each level of testing checked above, describe the method of validity testing and what it tests (*describe the steps—do not just name a method; what was tested, e.g., accuracy of data elements compared to authoritative source, relationship to another measure as expected; what statistical analysis was used)*

2017 Submission

We assessed face validity and construct validity for this measure.

Method of Assessing Face Validity: NCQA has identified and refined measure management into a standardized process called the HEDIS measure life cycle.

STEP 1: NCQA staff identifies areas of interest or gaps in care. Clinical expert panels (MAPs – whose members are authorities on clinical priorities for measurement) participate in this process. Once topics are identified, a literature review is conducted to find supporting documentation on their importance, scientific soundness, and feasibility. This information is gathered into a work-up format. Refer to What Makes a Measure "Desirable"? The work-up is vetted by NCQA's Measurement Advisory Panels (MAPs), the Technical Measurement Advisory Panel (TMAP) and the Committee on Performance Measurement (CPM) as well as other panels as necessary.

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

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

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

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

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

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

Method of testing construct validity: We tested for construct validity by exploring whether Colorectal Cancer Screening was correlated with Breast Cancer Screening. We hypothesized that organizations that perform well on Colorectal Cancer Screening should perform well on Breast Cancer Screening. To test these correlations, we used a Pearson correlation test. This test estimates the strength of the linear association between two continuous variables; the magnitude of correlation ranges from -1 to +1. A value of 1 indicates a perfect linear dependence in which increasing values on one variable is associated with increasing values of the second variable. A value of 0 indicates no linear association. A value of -1 indicates a perfect linear relationship in which increasing values of the first variable is associated with decreasing values of the second variable.

2011 Submission

NCQA tested the measure for face validity using a panel of stakeholders with specific expertise in measurement. This panel included representatives from key stake holder groups, including oncologists, family practitioners, health plans, state Medicaid agencies and researchers. Experts reviewed the results of the field test and assessed whether the results were consistent with expectation, whether the measure represented quality care, and whether we were measuring the most important aspects of care in this area.

In the pilot test, we explored periodicities associated with colorectal cancer screening, as long periodicities in light of average lengths of enrollment in MCOs can be a threat to validity. We examined whether the rates of screening would differ depending on the length of time an individual had been enrolled in the plan and found little effect as shown in Table 2. Although the rates increase a small amount each year in each plan, the relative rates of screening remain about the same. The sample sizes decline significantly with increased lengths of continuous enrollment; at 10 years, only two MCOs had enough data to estimate the rate.

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

2017 Submission

Results of face validity assessment:

Input from our multi-stakeholder measurement advisory panels and those submitting to public comment indicate the measure has face validity.

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

Table 1a. Correlations in Commercial Measures – 2016

	Pearson Correlation Coefficient
	Breast Cancer Screening
Colorectal Cancer Screening	0.711

Note: p<0.0001

Table 1b. Correlations in Medicare Measures – 2016

	Pearson Correlation Coefficient
	Breast Cancer Screening
Colorectal Cancer Screening	0.716

Note: p<0.0001

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?*)

2017 Submission

Interpretation of systematic assessment of face validity: These results indicate the technical expert panel showed good agreement that the measures as specified will accurately differentiate quality across providers. Our interpretation of these results is that this measure has sufficient face validity.

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

____

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 _risk factors

□ Stratification by _risk categories

□ Other,

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)*

2017 Submission

To demonstrate meaningful differences in performance, NCQA calculates an inter-quartile range (IQR) for each indicator. The IQR provides a measure of the dispersion of performance. The IQR can be interpreted as the difference between the 25th and 75th percentile on a measure. To determine if this difference is statistically significant, NCQA calculates an independent sample t-test of the performance difference between two randomly selected plans at the 25th and 75th percentile. The t-test method calculates a testing statistic based on the sample size, performance rate, and standardized error of each plan. The test statistic is then compared against a normal distribution. If the p value of the test statistic is less than 0.05, then the two plans' performance is significantly different from each other.

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*)

2017 Submission

HEDIS 2017 Variation in Performance across Health Plans

	Avg. EP	Avg.	SD	10th	25th	50th	75th	90th	IQR	p-value
Com.	8582	60.1	9.6	48.4	53.9	60.1	66.4	72.2	12.5	<0.001
Medicare	1330	67.7	12.4	50.8	60.9	69.9	76.4	81.0	15.5	<0.001

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

IQR: Interquartile range

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

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?*)

2017 Submission

The difference between the 25th and 75th percentile is statistically significant for both product lines. For commercial plans, there is a 12.5 percentage point gap between 25th and 75th percentile plans. This gap represents an average 1,073 more patients that have been screened for colorectal cancer compared to low performing plans (estimated from average health plan eligible population).

____

2b5. COMPARABILITY OF PERFORMANCE SCORES WHEN MORE THAN ONE SET OF SPECIFICATIONS

***If only one set of specifications, this section can be skipped*.**

Note*: This item is directed to measures that are risk-adjusted (with or without social risk factors)* ***OR*** *to measures with more than one set of specifications/instructions (e.g., one set of specifications for how to identify and compute the measure from medical record abstraction and a different set of specifications for claims or eMeasures). It does not apply to measures that use more than one source of data in one set of specifications/instructions (e.g., claims data to identify the denominator and medical record abstraction for the numerator).* ***Comparability is not required when comparing performance scores with and without social risk factors in the risk adjustment model. However, if comparability is not demonstrated for measures with more than one set of specifications/instructions, the different specifications (e.g., for medical records vs. claims) should be submitted as separate measures.***

2b5.1. Describe the method of testing conducted to compare performance scores for the same entities across the different data sources/specifications (*describe the steps—do not just name a method; what statistical analysis was used*)

2b5.2. What were the statistical results from testing comparability of performance scores for the same entities when using different data sources/specifications? (*e.g., correlation, rank order*)

2b5.3. What is your interpretation of the results in terms of the differences in performance measure scores for the same entities across the different data sources/specifications? (i*.e., what do the results mean and what are the norms for the test conducted*)

____

2b6. MISSING DATA ANALYSIS AND MINIMIZING BIAS

2b6.1. Describe the method of testing conducted to identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias (*describe the steps—do not just name a method; what statistical analysis was used*)

2017 Submission

This measure is collected with a complete sample.

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*)

2017 Submission

This measure is collected with a complete sample.

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*)

2017 Submission

This measure is collected with a complete sample.

3. Feasibility

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

3a. Byproduct of Care Processes

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

3a.1. Data Elements Generated as Byproduct of Care Processes.

Generated or collected by and used by healthcare personnel during the provision of care (e.g., blood pressure, lab value, diagnosis, depression score), Coded by someone other than person obtaining original information (e.g., DRG, ICD-9 codes on claims), Abstracted from a record by someone other than person obtaining original information (e.g., chart abstraction for quality measure or registry)

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**.

Some data elements are in defined fields in electronic sources

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).

To allow for widespread reporting across health plans and health care practices, this measure is collected through multiple data sources (administrative data, electronic clinical data, paper records, and registry). We anticipate as electronic health records become more widespread the reliance on paper record review will decrease.

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.

NCQA conducts an independent audit of all HEDIS collection and reporting processes, as well as an audit of the data which are manipulated by those processes, in order to verify that HEDIS specifications are met. NCQA has developed a precise, standardized methodology for verifying the integrity of HEDIS collection and calculation processes through a two-part program consisting of an overall information systems capabilities assessment followed by an evaluation of the MCO's ability to comply with HEDIS specifications. NCQA-certified auditors using standard audit methodologies will help enable purchasers to make more reliable "apples-to-apples" comparisons between health plans.

The HEDIS Compliance Audit addresses the following functions:

1) Information practices and control procedures

2) Sampling methods and procedures

3) Data integrity

4) Compliance with HEDIS specifications

5) Analytic file production

6) Reporting and documentation

In addition to the HEDIS audit, NCQA provides a system to allow "real-time" feedback from measure users. Our Policy Clarification Support System receives thousands of inquiries each year on over 100 measures. Through this system, NCQA responds immediately to questions and identifies possible errors or inconsistencies in the implementation of the measure. This system informs both annual updates to the measures as well as routine re-evaluation of measures. These processes include updating value sets and clarifying the specifications. Measures are re-evaluated on a periodic basis and when there is a significant change in evidence.

3c.2. Describe any fees, licensing, or other requirements to use any aspect of the measure as specified *(e.g., value/code set, risk model, programming code, algorithm)*.

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

4. Usability and Use

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

4a. Accountability and Transparency

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

4.1. Current and Planned Use

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

Specific Plan for Use	Current Use (for current use provide URL)
Public Reporting Quality Improvement (Internal to the specific organization)	Payment Program Medicare STARS https://www.medicare.gov/find-a-plan/questions/home.aspx California&acute;s Value Based Pay for Performance Program http://www.iha.org/our-work/accountability/value-based-p4p Quality Payment Program https://qpp.cms.gov Regulatory and Accreditation Programs Accreditation http://www.ncqa.org/Programs/Accreditation/Health-Plan-HP.aspx HEDIS ACO http://www.ncqa.org/Programs/Accreditation/AccountableCareOrganizationACO.aspx Quality Improvement (external benchmarking to organizations) Annual State of Health Care Quality http://www.ncqa.org/report-cards/health-plans/state-of-health-care-quality Quality Compass http://www.ncqa.org/hedis-quality-measurement/quality-measurement-products/quality-compass

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

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

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

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

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

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

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

NCQA ACCOUNTABLE CARE ORGANIZATION ACCREDITATION: This measure is used in NCQA's ACO Accreditation program, that helps health care organizations demonstrate their ability to improve quality, reduce costs and coordinate patient care. ACO standards and guidelines incorporate whole-person care coordination throughout the health care system.

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

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

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

N/A

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

N/A

4a2.1.1. Describe how performance results, data, and assistance with interpretation have been provided to those being measured or other users during development or implementation.

How many and which types of measured entities and/or others were included? If only a sample of measured entities were included, describe the full population and how the sample was selected.

Health plans that report HEDIS calculate their rates and know their performance when submitting to NCQA. NCQA publicly reports rates across all plans and also creates benchmarks in order to help plans understand how they perform relative to other plans. Public reporting and benchmarking are effective quality improvement methods.

4a2.1.2. Describe the process(es) involved, including when/how often results were provided, what data were provided, what educational/explanatory efforts were made, etc.

NCQA publishes HEDIS results annually in our Quality Compass tool. NCQA also presents data at various conferences and webinars. For example, at the annual HEDIS Update and Best Practices Conference, NCQA presents results from all new measures' first year of implementation or analyses from measures that have changed significantly. NCQA also regularly provides technical assistance on measures through its Policy Clarification Support System, as described in Section 3c.1.

4a2.2.1. Summarize the feedback on measure performance and implementation from the measured entities and others described in 4d.1.

Describe how feedback was obtained.

NCQA measures are evaluated regularly using a consensus-based process to consider input from multiple stakeholders, including but not limited to entities being measured. We use several methods to obtain input, including vetting of the measure with several multi-stakeholder advisory panels, public comment posting, and review of questions submitted to the Policy Clarification Support System. This information enables NCQA to comprehensively assess a measure's adherence to the HEDIS Desirable Attributes of Relevance, Scientific Soundness and Feasibility.

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

Questions received through the Policy Clarification Support system have generally centered around clarification on whether certain notation in medical record documentation is sufficient to meet measure criteria. Other questions have sought clarification about the screening methods that satisfy the measure numerator. During a recent public comment session, a majority of comments from measured entities supported updates to the measure to align with the latest clinical recommendations.

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

This measure has been deemed a priority measure by NCQA and other entities, as illustrated by its use in programs such as the Medicare Advantage Star Rating program.

4a2.3. Describe how the feedback described in 4a2.2.1 has been considered when developing or revising the measure specifications or implementation, including whether the measure was modified and why or why not.

During the measure's last major update, feedback obtained through the mechanisms described in 4a2.2.1 informed how we revised the measure to include new screening methods recommended by the U.S. Preventive Services Task Force and other major clinical guideline organizations.

Improvement

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

4b1. Refer to data provided in 1b but do not repeat here. Discuss any progress on improvement (trends in performance results, number and percentage of people receiving high-quality healthcare; Geographic area and number and percentage of accountable entities and patients included.)

If no improvement was demonstrated, what are the reasons? If not in use for performance improvement at the time of initial endorsement, provide a credible rationale that describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.

From 2014 to 2016, performance rates for this measure have been generally stable or shown slight improvement. In 2016, commercial plans on average performance rate of 60 percent, and Medicare plans had an average rate of 68 percent. There continues to be significant variation between the 10th and 90th percentiles, suggesting room for improvement. In 2016, commercial plans in the 10th percentile had a rate of 48 percent, compared to 72 percent among plans in the 90th percentile. For Medicare, plans in the 10th percentile had a rate of 51 percent compared to 81 percent among plans in the 90th percentile.

Given the new US Preventive Services Task Force guidelines for colorectal cancer screening and our recent changes to this measure, we may see performance improvement in the coming years. In 2016, two additional screening methods were added to the guideline and measure. The addition of more screening options may help patients feel more comfortable with the screening process, and therefore increase the number of patients who choose to be screened for colorectal cancer.

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.

There were no identified unintended consequences for this measure during testing or since implementation.

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

There were no identified unexpected benefits for this measure during testing or since implementation.

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)

0658 : Appropriate Follow-Up Interval for Normal Colonoscopy in Average Risk Patients

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

Colorectal Cancer Screening – Minnesota Community Measurement

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.

Minnesota Community Measurement: These measures are harmonized but intended for different levels of accountability. --Both measures exclude patients who have had a total colectomy, a history of colorectal cancer, or who have been in hospice care. --Both measures include the same screening methods and intervals. --The Minnesota Community Measurement quality measure is intended for use at the clinician or practice-level, whereas NQF#0034 is intended for use at the health plan level. American Gastroenterological Association: These measures have different areas of focus and are harmonized where appropriate. --The American Gastroenterological Association measure focuses on only one of the available screening methods: colonoscopy. The measure assesses whether patients who have had a colonoscopy also have a recommended follow-up interval of 10 years documented in their colonoscopy report.

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.)
Not applicable.

Appendix

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

No appendix Attachment:

Contact Information

Co.1 Measure Steward (Intellectual Property Owner): National Committee for Quality Assurance

Co.2 Point of Contact: Bob, Rehm, nqf@ncqa.org, 202-955-1728-

Co.3 Measure Developer if different from Measure Steward: National Committee for Quality Assurance

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

Additional Information

Ad.1 Workgroup/Expert Panel involved in measure development

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

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

2008 Colorectal Cancer Measurement Advisory Panel members:

Joel V. Brill, Predictive Health, LLC

Durado Brooks, American Cancer Society

Robert Fletcher, Harvard Medical School

William Lawrence, AHRQ Center for Outcomes and Effectiveness

T.R. Levin, Kaiser Permanente

Michael Pignone, UNC Hospital

Evelyn Whitlock

2016 Colorectal Cancer Screening Measurement Advisory Panel members:

Matthew Barish, MD FACR, Stony Brook University Hospital

Linda Berthold, PhD, Central California Alliance for Health

Durado Brooks, MD MPH, American Cancer Society

Joseph Chin, MD MS, Centers for Medicare and Medicaid Services

T.R. Levin, MD, Kaiser Permanente Northern California

Steven Phillips, MD CMD, Sierra Health Services Inc

Tim Wilt, MD MPH, VA Medical Center Minneapolis

Ann Zauber, PhD, Memorial Sloan Kettering Cancer Center

2016 Geriatric Measurement Advisory Panel members:

Wade Aubry, UCSF Institute for Health Policy Studies

Arlene Bierman, AHRQ

Patricia A. Bomba, MD FACP, Excellus BlueCross BlueShield

Jennie Chin Hansen, RN, American Geriatrics Society

Joyce Dubow, Public Member/Consumer Advocate

Peter Hollman, Brown University

Jeffrey Kelman, MMSc, MD, Centers for Medicare & Medicaid Services

Steven Phillips, MD, CMD, Geriatric Specialty Care

Eric G. Tangalos, MD, FACP, AGSF, CMD, Mayo Clinic

Dirk Wales, MD, PsyD, Cigna HealthSpring

Joan Weiss, PhD, RN, CRNP, U.S. Department of Health and Human Services

Neil Wenger, MD, UCLA Division of Medicine

2016 Committee on Performance Measurement members:

Bruce Bagley, MD, American Medical Association

Andrew Baskin, MD, Aetna

Jonathan D. Darer, MD, MPH, Medicalis

Helen Darling, National Quality Forum

Foster Gesten, MD, FACP, New York State Department of Health

Kate Goodrich, MD, MHS, Centers for Medicare and Medicaid Services

David Grossman, MD, MPH, Group Health Physicians

Christine S. Hunter, MD (Co-chair), US Office of Personnel Management

Jeffrey Kelman, MMSc, MD, United States Department of Health and Human Services (DHHS)

Nancy Lane, PhD, Vanderbilt University Medical Center

Bernadette Loftus, MD, The Permanente Medical Group

Adrienne Mims, MD, MPH, Alliant Quality

Amanda Parsons, MD, MBA, Montefiore Health System

J. Brent Pawlecki, MD, MMM, The Goodyear Tire & Rubber Company

Susan Reinhard, PhD, RN, AARP Public Policy Institute

Eric C. Schneider, MD, MSc, FACP (Co-chair), The Commonwealth Fund

Marcus Thygeson, MD, MPH, Blue Shield of California

JoAnn Volk, MA, Georgetown University Center on Health Insurance

Reforms

2016 Technical Measurement Advisory Panel members:

Andy Amster, MSPH, Kaiser Permanente

Jennifer Brudnicki, MBA, Geisinger Health Plan

Lindsay Cogan, PhD, MS, New York State Department of Health

Kathy Coltin, MPH, Independent Consultant

Mike Farina, MVP Healthcare

Marissa Finn, MBA, CIGNA HealthCare

Scott Fox, MS, Med,Independence Blue Cross

Carlos Hernandez, CenCal Health

Harmon Jordan, ScD, RTI International

Virginia Raney

Lynne Rothney-Kozlak, MPH, Rothney-Kozlak Consulting, LLC

Laurie Spoll, Aetna

Natan Szapiro, Independent Consultant

Measure Developer/Steward Updates and Ongoing Maintenance

Ad.2 Year the measure was first released: 2003

Ad.3 Month and Year of most recent revision: 10, 2016

Ad.4 What is your frequency for review/update of this measure? Approximately every 3 years, sooner if the clinical guidelines have changed significantly.

Ad.5 When is the next scheduled review/update for this measure? 2018

Ad.6 Copyright statement: © 2003 by the National Committee for Quality Assurance

1100 13th Street, NW, 3rd floor

Washington, DC 20005

Ad.7 Disclaimers: These performance measures are not clinical guidelines and do not establish a standard of medical care, and have not been tested for all potential applications. THE MEASURES AND SPECIFICATIONS ARE PROVIDED "AS IS" WITHOUT WARRANTY OF ANY KIND.

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

These performance measures were developed and are owned by NCQA. They are not clinical guidelines and do not establish a standard of medical care. NCQA makes no representations, warranties, or endorsement about the quality of any organization or physician that uses or reports performance measures, and NCQA has no liability to anyone who relies on such measures. NCQA holds a copyright in these measures and can rescind or alter these measures at any time. Users of the measures shall not have the right to alter, enhance, or otherwise modify the measures, and shall not disassemble, recompile, or reverse engineer the source code or object code relating to the measures. Anyone desiring to use or reproduce the measures without modification for a noncommercial purpose may do so without obtaining approval from NCQA. All commercial uses must be approved by NCQA and are subject to a license at the discretion of NCQA. © 2017 by the National Committee for Quality Assurance