NATIONAL QUALITY FORUM

MEASURE WORKSHEET

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

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

Brief Measure Information

NQF #: 0658

Corresponding Measures:

De.2. Measure Title: Appropriate Follow-Up Interval for Normal Colonoscopy in Average Risk Patients

Co.1.1. Measure Steward: American Gastroenterological Association

De.3. Brief Description of Measure: Percentage of patients aged 50 years to 75 years receiving a screening colonoscopy without biopsy or polypectomy who had a recommended follow-up interval of at least 10 years for repeat colonoscopy documented in their colonoscopy report.

1b.1. Developer Rationale: Guideline recommendations support screening colonoscopy at 10 year intervals, for average risk patients. Non-adherence to guideline recommendations increases patients to unnecessary risk via procedural harms and complications. Colonoscopy screening at more frequent intervals also contributes to increased costs to patients and insurers.

In the average-risk population, colonoscopy screening is recommended in all current guidelines at 10-year intervals. Inappropriate interval recommendations can result in overuse of resources and can lead to significant patient harm. Performing colonoscopy too often not only increases patients' exposure to procedural harm, but also drains resources that could be more effectively used to adequately screen those in need (Lieberman et al, 2008). The most common serious complication of colonoscopy is post-polypectomy bleeding (Levin et al, 2008).

Variations in the recommended time interval between colonoscopies exist for patients with normal colonoscopy findings. In a 2006 study of 1282 colonoscopy reports, recommendations were consistent with contemporaneous guidelines in only 39.2% of cases and with current guidelines in 36.7% of cases. Further, the adjusted mean number of years in which repeat colonoscopy was recommended was 7.8 years following normal colonoscopy (Krist et al, 2007)

S.4. Numerator Statement: Patients who had a recommended follow-up interval of at least 10 years for repeat colonoscopy documented in their colonoscopy report

S.6. Denominator Statement: All patients aged 50 years to 75 years and receiving screening a screening colonoscopy without biopsy or polypectomy

S.8. Denominator Exclusions: Documentation of medical reason(s) for not recommending at least a 10 year follow-up interval (eg, inadequate prep,familial or personal history of colonic polyps, patient had no adenoma and age is >= 66 years old, or life expectancy < 10 years, other medical reasons)

De.1. Measure Type: Process

S.17. Data Source: Claims, Electronic Health Data, Electronic Health Records, Other, Registry Data

S.20. Level of Analysis: Clinician : Individual

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**

* Colorectal cancer screening for the age range indicated is grade A, but the 10 year interval is not graded

Summary of prior review in [2013]

- The developer notes there is a significantamount of evidence to support this measure focus.
- There was discussion by the Committee on whether the 10-year interval specified in this measure is based on evidence or consensus, but did not reach a conclusion. Most polyps > 1 cm in diameter appear to grow for 5-10 years before becoming colorectal cancer. Usefulness of an interval beyond 10 years has not been studied.
- Committee members noted in 2013 that prospective studies have demonstrated that very few patients (< 3%) have advanced adenomas when colonoscopy is repeated 5 years after a normal screening colonoscopy.
- The developer attested in 2013 that there were no changes in the evidence since the measure was evaluated in 2011

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:** In 2017 the United States Task Force (USMSTF) guidance recommended colonoscopy every 10 years as a tier 1 recommendation which is a strong recommendation with moderate quality of evidence.

Exception to evidence
N/A

Questions for the Committee:

- *The developer cited the 2016 USPSTF guideline and the 2017 USMSTF guidline. Does the Committee agree these support the measure focus? What about the specific requirement in the measure related to follow-up interval of at least 10 years?*

Guidance from the Evidence Algorithm

Measure a health outcome (Box 1) No → Assess performance of intermediate outcome, process, or structure(Box 3) Yes → Empirical evidence without SR or QQC (Box 4) yes → Grade for evidence (Box 6) Yes → Moderate

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

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

Maintenance measures – increased emphasis on gap and variation

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

There is variation in the recommendations made to patients that differ from existing guidelines. Patients may receive a recommendation "consistent with contemporaneous guidelines in only 39.2% of cases and with current guidelines in 36.7%." These variations err on the side of increased frequency of procedures, leading to overuse of resources and potential for patient harm.

- The developer reported variations in the recommended time interval between colonoscopies exist for patients with normal colonoscopy findings. A 2006 study of 1,282 colonoscopy reports, recommendations were consistent with contemporaneous guidelines in only 39.2% of cases and with current guidelines in 36.7% of cases. Further, the adjusted mean number of years in which repeat colonoscopy was recommended was 7.8 years following normal colonoscopy (Krist et al, 2007).
- The developer provided performance data from 2016-2018 is provided at the individual physician level with the mean, standard deviation, minimum and maximum performance as well as the interquartile range (IQR).
    - Individual physician
        - Measurement Year: 2016; 2017; 2018
        - Number of physicians: 3,136; 3,618; 3,747
        - Mean: 85.12; 85.63; 85.43
        - Std Dev: 23.71; 23.32; 23.21
        - Min: 0; 0; 0
        - Max: 100; 100; 100
        - IQR: 15.36; 14.99; 15.67

Disparities
The developer provides disparity data by age, ethnicity, and age by individual physician.

Questions for the Committee:

- *Does the Committee feel there is a performance gap and so opportunity for improvement?*

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

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

1a. Evidence to Support Measure Focus

- This is an existing measure with updated evidence in support of the measure.
- Data appears directly relevant. I am not aware of any new evidence that changes the evidence base regarding the importance of this screening procedure.
- There is ample evidence to support colorectal cancer screenign for the indicatad age group. I do have questions about the "at least 10 years" phrasing. A clinician whose patient waits 25 years for a f/u would be "in complaince" with the measure though not necessarily from a prevention focus. Has an upper limit ever been discussed / proposed? That is, "10-15 years", say.
- The evidence is appropriate.
- The evidence cited is from USPSTF and USMSTF guidelines recommending colorectal screening at certain intervals, which is not the same thing as the recommended follow-up interval being at least 10 years. And the existing guidelines are mostly not based on RCTs. In addition, the USPSTF now recommends other screening methods such as FIT and FIT-DNA, for which the recommended follow-up intervals are 1 and 3 years respectively. Furthermore, the recommended interval is not what matters; rather it is the actual interval, but this is not addressed in the proposed measure. Evidence rating: Low
- The evidence cited is from USPSTF and USMSTF guidelines recommending colorectal screening at certain intervals, which is not the same thing as the recommended follow-up interval being at least 10 years. And the existing guidelines are mostly not based on RCTs. In addition, the USPSTF now recommends other screening methods such as FIT and FIT-DNA, for which the recommended follow-up intervals are 1 and 3 years respectively. Furthermore, the recommended interval is not what matters; rather it is the actual interval, but this is not addressed in the proposed measure.
- There is empirical data to support this process measure. However, it should be noted that the American Cancer Society has updated the colonoscopy cancer screening guidelines to >45 years of age--not >50.

1b. Performance Gap

- There is a performance gap
- There still appear to be opportunities to improve adherence to the recommendations; performance overall is less than optimal. Population subgroups were provided and demonstrate some relevant disparities.
- Opportunities for improvement have been documented. Analyses stratified by race were provided: racial and ethnic minorities patients seem more like to have an appropriate f/u documented..
- I agree with the moderate rating.
- Developers cite a 2007 study indicating that “Patients may receive a recommendation ‘consistent with contemporaneous guidelines in only 39.2% of cases and with current guidelines in 36.7%,’” but their own data suggests that this proportion is roughly 15% at the individual physician level and 12% at the practice/group level. The developers offer no data at all regarding disparities. Opportunity for improvement rating: Low
- Developers cite a 2007 study indicating that “Patients may receive a recommendation ‘consistent with contemporaneous guidelines in only 39.2% of cases and with current guidelines in 36.7%,’” but their own data suggests that this proportion is roughly 15% at the individual physician level and 12% at the practice/group level. The developers offer no data at all regarding disparities.
- Opportunities for improvement on this measure is moderate given the national performance gap, population subgroups and disparities in care.

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability: Specifications and Testing

2b. Validity: Testing; Exclusions; Risk-Adjustment; Meaningful Differences; Comparability; Missing Data

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.

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

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

Reliability

- The developer provided updated score-level reliability testing using a beta-binomial model and measuring the ratio of signal to noise at the physician level.
- Two data sources were used for testing: Data 1 = registry data from the Physician Quality Reporting System (PQRS), provided by the Centers for Medicare & Medicaid Services (CMS) (2016 data). Data 2 = GI Quality Improvement Consortium, Ltd (GIQuIC), a non-profit educational and scientific organization established by the American College of Gastroenterology (ACG) and the American Society for Gastrointestinal Endoscopy (ASGE) that is an approved Qualified Clinical Data Registry (QCDR) (2016 data).
- The developer states that, overall, the data suggest that for physicians with an average or greater number of events the measure has high reliability.
    - The developer reports a reliability statistic of 0.90 for the CMS data set; 237 physicians had all the required data elements and met the minimum number of quality reporting events (10).
    - The developer reports a reliability statistic of 0.94 for the GIQuIC data set; 2,666 physicians had all the required data elements and met the minimum number of quality reporting events.

Validity

- The developer conducted construct validity, using Colorectal Cancer Screening (PQRS #113) for correlation analysis due to the similarities in patient population and domain. It hypothesized a positive association between patients receiving a screening colonoscopy (PQRS #113) and those who had documentation of appropriate recommended follow-up interval of at least 10 years for repeat colonoscopy (this measure). The developer could only provide correlation analysis for the CMS data set (237 physicians). For this analysis, the coefficient was 0.20 and p-value = 0.007. The developer states this result is a moderate positive correlation.
- Face validity had been previously performed by the developer.

Questions for the Committee regarding reliability:

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

Questions for the Committee regarding validity:

- Do you have any concerns regarding the validity of the measure (e.g., exclusions, risk-adjustment approach, etc.)?
- The staff is satisfied with the validity analyses for the measure. Does the Committee think there is a need to discuss and/or vote on validity?

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

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

Scientific Acceptability Evaluation

Scientific Acceptability: Preliminary Analysis Form

Measure Number: 0658

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

Type of measure:

☒ **Process** ☐ **Process: Appropriate Use** ☐ **Structure** ☐ **Efficiency** ☐ **Cost/Resource Use**
☐ **Outcome** ☐ **Outcome: PRO-PM** ☐ **Outcome: Intermediate Clinical Outcome** ☐ **Composite**

Data Source:

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

Level of Analysis:

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

Measure is:

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

RELIABILITY: SPECIFICATIONS

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

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

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

2. **Briefly summarize any concerns about the measure specifications.**

   The measure does not specify the risk level of patients. However, the guideline recommendation which is used to support the measure is based on patients with average risk and does not address patients with low and /or high risk.

RELIABILITY: TESTING

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

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

   ☐ **Yes** ☐ **No**

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

   Submission document: Testing attachment, section 2a2.2

   - Reliability was assessed using a beta-binomial model and measuring the ratio of signal to noise. For this measure, the signal is the proportion of the variability in measured performance that can be explained by real differences in physician performance. Reliability is the ratio of the physician-to-physician variance divided by the sum of the physician-to-physician variance plus the error variance specific to a physician.
   - Reliability testing was completed using two different data sources: Data 1 = registry data from the Physician Quality Reporting System (PQRS), provided by the Centers for Medicare & Medicaid Services (CMS) (2016 data). Data 2 = GI Quality Improvement Consortium, Ltd (GIQuIC), a non-profit educational and scientific organization established by the American College of Gastroenterology (ACG) and the American Society for Gastrointestinal

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

   Submission document: Testing attachment, section 2a2.3

   - 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 physician performance.
     - The developer reports a reliability statistic of 0.90 for the CMS data set; 237 physicians had all the required data elements and met the minimum number of quality reporting events (10).
     - The developer reports a reliability statistic of 0.94 for the GIQuIC data set; 2,666 physicians had all the required data elements and met the minimum number of quality reporting events (10).

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

   Submission document: Testing attachment, section 2a2.2

☒ **Yes**

☐ **No**

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

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

   Submission document: Testing attachment, section 2a2.2

   ☒ **Yes**

   ☐ **No**

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

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

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

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

    ☐ **Low** (NOTE: Should rate LOW if you believe specifications are NOT precise, unambiguous, and complete or if testing methods/results are not adequate)

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

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

    - The beta-binomial method used is appropriate and common for these types of measures. The reliability statistic reported is considered within the literature to be high.

VALIDITY: ASSESSMENT OF THREATS TO VALIDITY

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

    Submission document: Testing attachment, section 2b2.

    - The developer states that exceptions/exclusions were determined based on reported characteristics of the endoscopy. Some of the possible reasons for a denominator exception could be: inadequate bowel prep; incomplete colon examination; above average patient risk; complications arising during colonoscopy.
        - For the CMS data set, the exception rate was 10%, with a range of 0-73%.
        - For the GIQuIC data set, the exception rate was 11%, with a range of 0-64%.
    - The developer states that these rates are "fairly consistent" with research that finds up to 20-25% of colonoscopies are reported to have an inadequate bowel preparation (one of the exceptions) making it appropriate for a patient to have a follow-up interval of less than 10 years for repeat colonoscopy.

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

    Submission document: Testing attachment, section 2b4.

    - For the CMS data set, the interquartile range is 0.09 (1.0-0.91). The mean performance rate is 0.93 the median performance rate is 1.00 and the mode is 1.0. The standard deviation is 0.15. The range of the performance rate is 0.95, with a minimum rate of 0.05 and a maximum rate of 1.0.
    - For the GIQuIC data set, the interquartile rate is 0.12 (0.98-0.86). The mean performance rate is 0.88 the median performance rate is 0.94 and the mode is 1.0. The standard deviation is 0.17. The range of the performance rate is 0.99, with a minimum rate of 0.01 and a maximum rate of 1.0. The interquartile range is 0.12 (.98–0.86).
    - The developer concludes these results demonstrate meaningful differences in performance.

14. **Please describe any concerns you have regarding comparability of results if multiple data sources or methods are specified.**
Submission document: Testing attachment, section 2b5.
Not applicable

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

Submission document: Testing attachment, section 2b6.

- The developer noted that the CMS dataset does not contain missing data. Therefore, missing data tests were not performed. The developer did note that missing data may have been rejected when submitted to CMS, in which case those values would not be counted towards measure performance. The developer asserts there is no indication that this missing data was systematic. Missing data as it relates to the GIQuIC data set was not discussed.

16. **Risk Adjustment**

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

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

☐ Yes ☐ No ☒ Not applicable

16c. **Social risk adjustment:**

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

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

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

16d.**Risk adjustment summary:**

16d.1 All of the risk-adjustment variables present at the start of care? ☐ Yes ☒ No
16d.2 If factors not present at the start of care, do you agree with the rationale provided for inclusion? ☐ Yes ☒ No
16d.3 Is the risk adjustment approach appropriately developed and assessed? ☐ Yes ☒ No
16d.4 Do analyses indicate acceptable results (e.g., acceptable discrimination and calibration) ☐ Yes ☒ No
16d.5.Appropriate risk-adjustment strategy included in the measure? ☐ Yes ☒ No

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

The measure is not risk adjusted**.**

VALIDITY: TESTING

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

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

☒ **Face validity**

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

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

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

Submission document: **Testing attachment, section 2b2.2**

- The developer conducted construct validity, using Colorectal Cancer Screening (PQRS #113) for correlation analysis due to the similarities in patient population and domain. It hypothesized a positive association between patients receiving a screening colonoscopy (PQRS #113) and those who had documentation of appropriate recommended follow-up interval of at least 10 years for repeat colonoscopy (this measure). The developer could only provide correlation analysis for the CMS data set (237 physicians).

- Face validity had been previously performed by the developer.

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

    Submission document: Testing attachment, section 2b2.3

    - The construct validity correlating the measure to Colorectal Cancer Screening yielded a coefficient of 0.20; p-value = 0.007. The developer states this result is a moderate positive correlation (citing to Shortell).

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

    Submission document: Testing attachment, section 2b1.

    ☒ **Yes**

    ☐ **No**

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

22. **Was the method described and appropriate for assessing the accuracy of ALL critical data elements?**

    NOTE that data element validation from the literature is acceptable.

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

    ☐ **Yes**

    ☐ **No**

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

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

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

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

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

    ☐ **Insufficient** (NOTE: For instrument-based measures and some composite measures, testing at both the score level and the data element level is required; if not conducted, should rate as INSUFFICIENT.)

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

The developer provided construct validity testing (correlating to a general colorectal screening measure) at the score level, but only for the CMS data set (237 physicians, and the resulting coefficient was 0.20. Meaningful differences were examined. The developer notes that the data set from CMS did not contain any missing data, and it hypothesized that submissions with missing data were not accepted and, regardless, were unlikely to be systematic.

ADDITIONAL RECOMMENDATIONS

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

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

2a1. Reliability-Specifications

- I don't think we need to discuss reliablity
- Data elements appear clearly defined. No concerns re: specs.
- No concerns
- Reliability is appropriate.
- Perhaps recommending the interval to the next screening is a regular part of colonoscopy reports, but if not, I would imagine that finding the recommended interval in a free-form report would be difficult.
- Perhaps recommending the interval to the next screening is a regular part of colonoscopy reports, but if not, I would imagine that finding the recommended interval in a free-form report would be difficult.
- Do the specifications need to be updated to reflect the new denominator: All patients aged 45 years to 75 years and receiving screening a screening colonoscopy without biopsy or polypectomy?

2a2. Reliability – Testing, any concerns?

- No
- No
- No
- I agree with high rating.
- The beta-binomial methods was appropriately used, with acceptable results for the average physician and practice/group. However, the developer allows the measure to be reported for physicians and or practices/groups with as few as 10 events (although the measure specifications do not mention this limitation). Even if the reliability for the average physician is acceptable, it's hard to imagine that any measure would have reasonable reliability based on so few events. Assuming binomial distributions, with n = 10 and p = 0.8, the s.d. is 0.126; with p = 0.5, the s.d. is 0.158. With n = 100, the s.d.s are 0.040 and 0.050 respectively. Not sure what to make about comment about level of testing on p. 6 of the measure worksheet. Reliability rating: Moderate
- The beta-binomial methods was appropriately used, with acceptable results for the average physician and practice/group. However, the developer allows the measure to be reported for physicians and or practices/groups with as few as 10 events (although the measure specifications do not mention this limitation). Even if the reliability for the average physician is acceptable, it's hard to imagine that any measure would have reasonable reliability based on so few events. Assuming binomial distributions, with n = 10 and p = 0.8, the s.d. is 0.126; with p = 0.5, the s.d. is 0.158. With n = 100, the s.d.s are 0.040 and 0.050 respectively. Not sure what to make about comment about level of testing on p. 6 of the measure worksheet.
- High rate of reliability

2b1. Validity -Testing, any concerns?

- No
- No
- No concerns
- Validity is appropriate.
- None.
- None.
- High rate of validity

2b4-7; 2b2-3. Threats to Validity

- No concerns
- No concerns about ability to determine meaningful differences. No concern regarding missing data
- None of note
- No concerns noted.
- None.
- None.
- N/A
- No concerns
- Exclusions appear consistent with evidence. No concerns re: risk adjustment.
- Colonoscopy rates have been a lonstanding health equity and public health / prevention issue and there has been significant sucess in closing gaps. Given the historic inequity in screening practices, I was surprised no conceptual description of the impact of social factors. The SDOH do play a role here and it would be worth assessing, particularly inequiteis by socioeconomic status.
- No concerns noted.
- Reporting the mean and standard deviation does not address the question of whether the proposed measure is able to statistically identify meaningful differences.
- NA
- No risk adjustment noted--which could be a concern

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.

- Currently used in professional certification programs, payment programs and for public reporting. Measure is reported via claims and registry data, which increases measure reporting feasibility. Data can be collected electronically via endowriter, an automated endoscopy record system (not an EHR/EMR) or manually via a web portal.

Questions for the Committee:

- None

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

Committee Pre-evaluation Comments:
Criteria 3: Feasibility

- Currently in use.
- Appears feasible, currently integrated into EHRs
- No concerns
- I agree with moderate rating.
- The measure is currently in use with no apparent problems.
- The measure is currently in use with no apparent problems.
- Feasible to collect data through registry, EHR, etc.

Criterion 4: Usability and Use

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

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

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

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

Current uses of the measure

Publicly reported? ☒ **Yes** ☐ **No**

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

OR

Planned use in an accountability program? ☐ **Yes** ☐ **No**

Accountability program details

- Quality Payment Program
    - This measure is currently publicly reported in the Quality Payment Program as a high-priority measure and has been reported in Physician Quality Reporting System (PQRS) since 2009. Multiple QCDRs facilitate participation in PQRS: Able Health, Academic Research for Clinical Outcomes (ARCO) in collaboration with ReportingMD, Inc; Citiustech, Inc.; Health-Advanta; Meditab Software, Inc.; Med-Xpress Registry; New Hampshire Colonoscopy Registry, Searfoss Consulting Group, Sovereign QCDR Registry.
- GI Quality Improvement Consortium, Ltd.
    - The GI Quality Improvement Consortium, Ltd. ("GIQuIC") is an educational and scientific 501(c)(3) organization established by gastroenterologists, physicians specializing in digestive disorders. GIQuIC is a joint initiative of the American College of Gastroenterology (ACG) and the American Society for Gastrointestinal Endoscopy (ASGE).
    - GIQuIC is a procedure-focused benchmarking registry using established quality indicators. The geographic area is the entire United States. GIQuIC registry participants have contributed real-time procedure related data from over 100,000 colonoscopies, not claims data, and the growth rate for the registry has increased to almost 2,000 new cases per week in recent months, with an accompanying surge in the growth of the number of practices involved in this quality improvement effort.
    - GIQuIC is a national registry that fosters the ability of endoscopists and endoscopy facilities to benchmark themselves, and provides impetus for quality improvement. Some 84 data fields for colonoscopy are collected and ten quality measures are benchmarked, including rate of cecal intubation, adenoma detection rate, prep assessment, and appropriate indications for procedure, among others. Currently, hundreds of physicians from endoscopy centers

nationwide have registered to participate in this ground-breaking initiative. http://giquic.gi.org/

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

Feedback on the measure by those being measured or others

- The measure has been implemented in the Quality Payment Program (QPP) as an individual measure for claims and registry reporting where feedback is provided via CMS QRUR reports. Measure is also implemented in multiple QCDRs where feedback is required quarterly.
- Feedback was received to include information about familial or personal history of colonic polyps as well as life expectancy of patient. Measure was modified to include in denominator exception in 2016.
- Measure was reviewed with ASC contractor, physician experts, and all three GI societies. Consensus was reached and measure was modified to include denominator exceptions. The developer provided examples in the measure exception language of instances that may constitute an exception and are intended to serve as a guide to clinicians.

Additional Feedback:

Measure has been implemented in the Quality Payment Program (QPP) as an individual measure for claims and registry reporting, feedback is provided via CMS QRUR reports. Measure is also implemented in multiple QCDRs where feedback is required quarterly. The developer reports that no feedback has been obtained by those beign measured or other users.

Questions for the Committee:

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

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

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

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

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

Improvement results

- The developer stated that "performance measurement serves as an important component in a quality improvement strategy but performance measurement alone will not achieve the desired goal of improving patient care. Measures can have their greatest effect when they are used judiciously and linked directly to operational steps that clinicians, patients, and health plans can apply in practice to improve care."
- The developer provides data from the literature, but no data from the measure per se. From the literature, a 2006 study of 1,282 colonoscopy reports, recommendations were consistent with contemporaneous guidelines in only 39.2% of cases and with current guidelines in 36.7% of cases.

Further, the adjusted mean number of years in which repeat colonoscopy was recommended was 7.8 years following normal colonoscopy (Krist et al, 2007). Therefore, opportunity for improvement exists.

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

Unexpected findings (positive or negative) during implementation

- The developer did not describe any unexpected findings during implementation.

Potential harms

- The developer was not aware of any unintended consequences related to this measurement.

Additional Feedback:

Measure has been implemented in the Quality Payment Program (QPP) as an individual measure for claims and registry reporting, feedback is provided via CMS QRUR reports. Measure is also implemented in multiple QCDRs where feedback is required quarterly. The developer reports that no feedback has been obtained by those beign measured or other users.

Questions for the Committee:

- Can the performance results be used to further the goal of high-quality, efficient healthcare?
- Is the Committee concerned about the relative lack of improvement during the 2016-2018 timeframe? (Is this measure, in effect, topping out?)
- Do the benefits of the measure outweigh any potential unintended consequences?

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

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

4a1. Use - Accountability and Transparency

- Currently used
- Currently reported as part of QPP/PQRS. Statement indicates that feedback from those being measured has been previously incorporated.
- This publicly reported measure is used in various programs. No concerns
- I agree with passing rating.
- The measure is currently in use and publicly reported with no apparent problems.
- The measure is currently in use and publicly reported with no apparent problems.
- This measure is a high priority measure in the QPP program.

4b1. Usability – Improvement

- I'm not concerned about it "topping" out and would encourage analysis by patient race/ethnicity, gender, socio/economic status as there may be disparities that could be addressed.
- Statement says that measure can be used for improvement in conjunction with operational changes at clinic level. No harms noted.
- I do not perceive any unintended harms
- I agree with moderate rating.
- NA
- The measure is currently in use and publicly reported with no apparent problems.
- High rate of usability

Criterion 5: Related and Competing Measures

Related or competing measures

0572 : Follow-up after initial diagnosis and treatment of colorectal cancer: colonoscopy

0659 : Colonoscopy Interval for Patients with a History of Adenomatous Polyps- Avoidance of Inappropriate Use

ASC-9: Appropriate Follow-up Interval for Normal Colonoscopy in Average Risk Patients - Telligen

ASC-10: Colonoscopy Interval for Patients with a History of Adenomatous Polyps – Avoidance of Inappropriate Use - Telligen

Harmonization

The list of measures above, includes several different populations and capture different elements in the numerator. The developer states that none of them are aiming to capture the same information as measure 0658.

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

- ASC-9 may be similar, but isn't NQF endorsed.
- No additional steps noted for harmonization. Related measures do not appear to substantially overlap.
- No concerns noted.
- NA
- NA
- Yes, appears to be competing measures and this measure is not completely harmonised with other colonoscopy cancer screening 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

Brief Measure Information

NQF #: 0658

Corresponding Measures:

De.2. Measure Title: Appropriate Follow-Up Interval for Normal Colonoscopy in Average Risk Patients

Co.1.1. Measure Steward: American Gastroenterological Association

De.3. Brief Description of Measure: Percentage of patients aged 50 years to 75 years receiving a screening colonoscopy without biopsy or polypectomy who had a recommended follow-up interval of at least 10 years for repeat colonoscopy documented in their colonoscopy report.

1b.1. Developer Rationale: Guideline recommendations support screening colonoscopy at 10 year intervals, for average risk patients. Non-adherence to guideline recommendations increases patients to unnecessary risk via procedural harms and complications. Colonoscopy screening at more frequent intervals also contributes to increased costs to patients and insurers.

In the average-risk population, colonoscopy screening is recommended in all current guidelines at 10-year intervals. Inappropriate interval recommendations can result in overuse of resources and can lead to significant patient harm. Performing colonoscopy too often not only increases patients' exposure to procedural harm, but also drains resources that could be more effectively used to adequately screen those in need (Lieberman et al, 2008). The most common serious complication of colonoscopy is post-polypectomy bleeding (Levin et al, 2008).

Variations in the recommended time interval between colonoscopies exist for patients with normal colonoscopy findings. In a 2006 study of 1282 colonoscopy reports, recommendations were consistent with contemporaneous guidelines in only 39.2% of cases and with current guidelines in 36.7% of cases. Further, the adjusted mean number of years in which repeat colonoscopy was recommended was 7.8 years following normal colonoscopy (Krist et al, 2007)

S.4. Numerator Statement: Patients who had a recommended follow-up interval of at least 10 years for repeat colonoscopy documented in their colonoscopy report

S.6. Denominator Statement: All patients aged 50 years to 75 years and receiving screening a screening colonoscopy without biopsy or polypectomy

S.8. Denominator Exclusions: Documentation of medical reason(s) for not recommending at least a 10 year follow-up interval (eg, inadequate prep,familial or personal history of colonic polyps, patient had no adenoma and age is >= 66 years old, or life expectancy < 10 years, other medical reasons)

De.1. Measure Type: Process

S.17. Data Source: Claims, Electronic Health Data, Electronic Health Records, Other, Registry Data

S.20. Level of Analysis: Clinician : Individual

IF Endorsement Maintenance – Original Endorsement Date: Jan 17, 2011 **Most Recent Endorsement Date:** Aug 14, 2013

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

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

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

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

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

1a. Evidence to Support the Measure Focus – See attached Evidence Submission Form

NQF_-658_Evidence_Form_07_16_12_final-636426432393177192.docx,AGA0658_Evidence_Attachment.docx

1a.1 <u>For Maintenance of Endorsement:</u> Is there new evidence about the measure since the last update/submission?

Do not remove any existing information. If there have been any changes to evidence, the Committee will consider the new evidence. Please use the most current version of the evidence attachment (v7.1). Please use red font to indicate updated evidence.

No

1a. Evidence (subcriterion 1a)

NATIONAL QUALITY FORUM—Evidence (subcriterion 1a)

Measure Number (*if previously endorsed*)**:** 0658

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

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

Date of Submission: 1/7/2020

Instructions

- ***Complete 1a.1 and 1a.2 for all measures. If instrument-based measure, complete 1a.3.***
- ***Complete EITHER 1a.2, 1a.3 or 1a.4 as applicable for the type of measure and evidence.***
- ***For composite performance measures:***
  - ***A separate evidence form is required for each component measure unless several components were studied together.***
  - ***If a component measure is submitted as an individual performance measure, attach the evidence form to the individual measure submission.***
- **All information needed to demonstrate meeting the evidence sub criterion (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.**

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

1a. Evidence to Support the Measure Focus

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

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

Notes

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

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

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

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

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

Outcome

☐ Outcome: Click here to name the health outcome

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

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

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

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

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

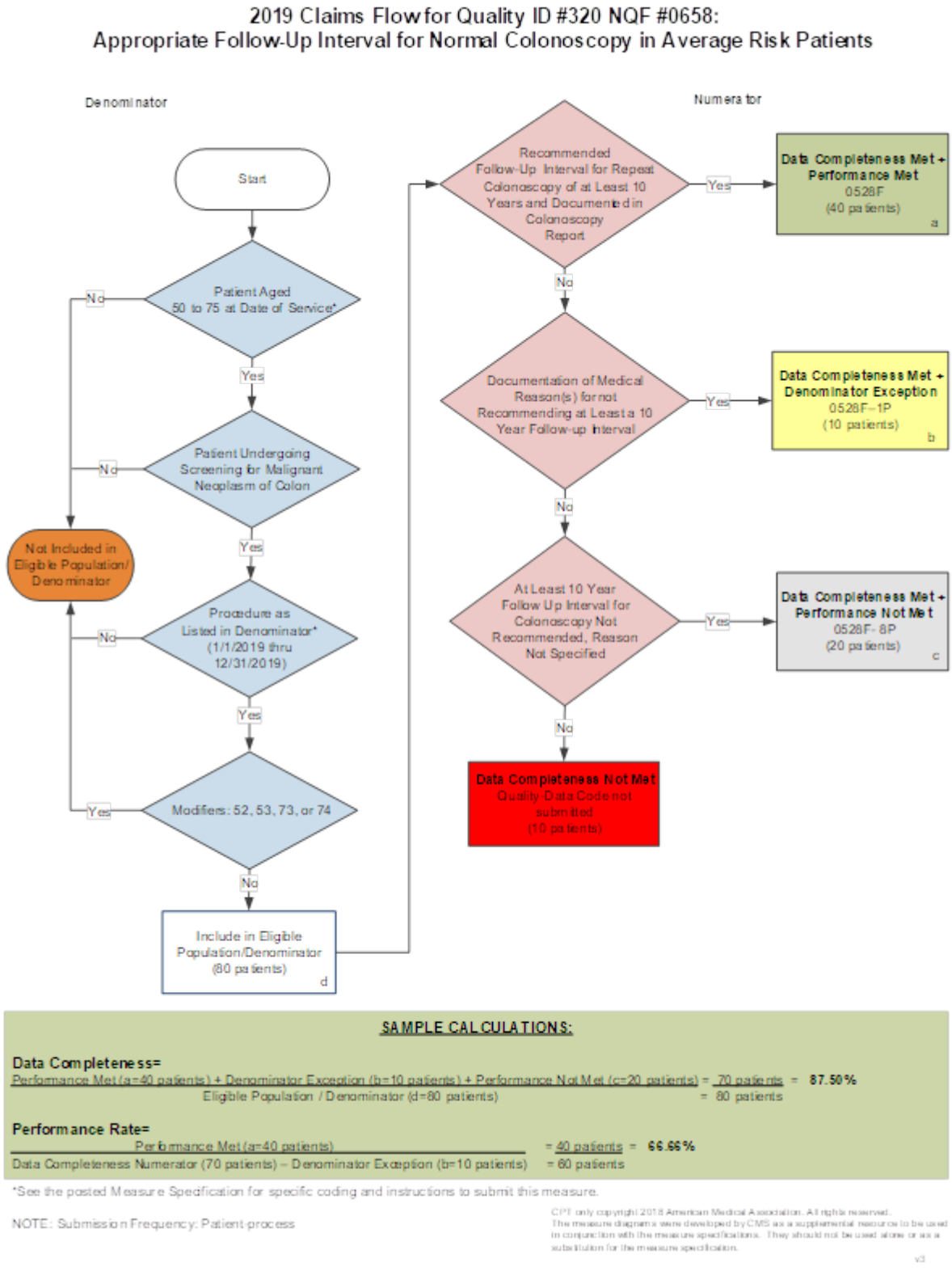
☐ Structure: Click here to name the structure

☐ Composite: Click here to name what is being measured

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

Logic Model provided by QMMS/CMS during measure maintenance cycle and performance data was provided by the GI Quality Improvement Consortium, Ltd (GIQuIC), a non-profit collaboration of the American College of Gastroenterology (ACG) and the American Society for Gastrointestinal Endoscopy (ASGE), which established

the GIQuIC clinical benchmarking registry, an approved Qualified Clinical Data Registry (QCDR). Performance from 2016-2018 is provided at the individual physician and practice/group level with the mean, standard deviation, minimum and maximum performance as well as the interquartile range (IQR).



SAMPLE CALCULATIONS:

Data Completeness=

Performance Met (a=40 patients) + Denominator Exception (b=10 patients) + Performance Not Met (c=20 patients) = 70 patients = **87.50%**
Eligible Population / Denominator (d=80 patients) = 80 patients

Performance Rate=

Performance Met (a=40 patients) = 40 patients = **66.66%**
Data Completeness Numerator (70 patients) – Denominator Exception (b=10 patients) = 60 patients

*See the posted Measure Specification for specific coding and instructions to submit this measure.

NOTE: Submission Frequency: Patient-process

The data provided to demonstrate a continued opportunity for improvement is from the GI Quality Improvement Consortium, Ltd (GIQuIC), a non-profit collaboration of the American College of Gastroenterology (ACG) and the American Society for Gastrointestinal Endoscopy (ASGE), which established the GIQuIC clinical benchmarking registry, an approved Qualified Clinical Data Registry (QCDR). Performance from 2016-2018 is provided at the individual physician and practice/group level with the mean, standard deviation, minimum and maximum performance as well as the interquartile range (IQR).

Individual physician
Measurement Year: 2016; 2017; 2018
Number of physicians: 3,136; 3,618; 3,747
Mean: 85.12; 85.63; 85.43
Std Dev: 23.71; 23.32; 23.21
Min: 0; 0; 0
Max: 100; 100; 100
IQR: 15.36; 14.99; 15.67

Practice/group
Measurement Year: 2016; 2017; 2018
Number of groups: 495; 581; 592
Mean: 87.79; 88.29; 87.71
Std Dev: 15.18; 15; 15.54
Min: 0; 0; 0
Max: 100; 100; 100
IQR: 11.72; 11.43; 12.20

The data provided to demonstrate a continued opportunity for improvement is from GI Quality Improvement Consortium, Ltd (GIQuIC), a non-profit collaboration of the American College of Gastroenterology (ACG) and the American Society for Gastrointestinal Endoscopy (ASGE), which established the GIQuIC clinical benchmarking registry, an approved Qualified Clinical Data Registry (QCDR). Overall performance on the measure from 2016-2018 stratified by age and race/ethnicity is provided with the mean, standard deviation, minimum and maximum performance at the individual physician and practice/group levels.

Individual Physician Level:
Age: 18-65
Measurement Year: 2016; 2017; 2018
Number of physicians: 3,142; 3,634; 3,763
Mean: 83.71; 84.18; 83.96
Std Dev: 24.37; 24.03; 23.97
Min: 0; 0; 0
Max: 100; 100; 100

Age: >65
Measurement Year: 2016; 2017; 2018

Number of physicians: 3,142; 3,634; 3,763
Mean: 100; 100; 100
Std Dev: 0; 0; 0
Min: 100; 100; 100
Max: 100; 100; 100

Race: White
Measurement Year: 2016; 2017; 2018
Number of physicians: 3,142; 3,634; 3,763
Mean: 85.08; 85.81; 85.48
Std Dev: 24.62; 23.61; 23.71
Min: 0; 0; 0
Max: 100; 100; 100

Race: Black
Measurement Year: 2016; 2017; 2018
Number of physicians: 3,142; 3,634; 3,763
Mean: 86.10; 87.26; 87.19
Std Dev: 27.01; 25.79; 25.57
Min: 0; 0; 0
Max: 100; 100; 100

Race: Asian
Measurement Year: 2016; 2017; 2018
Number of physicians: 3,142; 3,634; 3,763
Mean: 90.79; 89.88; 91.00
Std Dev: 24.96; 26.06; 28.37
Min: 0; 0; 0
Max: 100; 100; 100

Race: Other
Measurement Year: 2016; 2017; 2018
Number of physicians: 3,142; 3,634; 3,763
Mean: 86.26; 88.56; 87.33
Std Dev: 28.95; 26.84; 28.37
Min: 0; 0; 0
Max: 100; 100; 100

Race: Unknown
Measurement Year: 2016; 2017; 2018
Number of physicians: 3,142; 3,634; 3,763

Mean: 86.43; 87.32; 87.08
Std Dev: 25.54; 24.65; 24.89
Min: 0; 0; 0
Max: 100; 100; 100

Ethnicity: Hispanic Latino
Measurement Year: 2016; 2017; 2018
Number of physicians: 3,142; 3,634; 3,763
Mean: 89.19; 88.61; 88.95
Std Dev: 25.58; 26.90; 26.29
Min: 0; 0; 0
Max: 100; 100; 100

Ethnicity: Non-Hispanic Latino
Measurement Year: 2016; 2017; 2018
Number of physicians: 3,142; 3,634; 3,763
Mean: 85.29; 85.90; 85.11
Std Dev: 25.02; 24.20; 24.58
Min: 0; 0; 0
Max: 100; 100; 100

Ethnicity: Unknown
Measurement Year: 2016; 2017; 2018
Number of physicians: 3,142; 3,634; 3,763
Mean: 86.00; 86.45; 86.22
Std Dev: 24.50; 24.55; 24.47
Min: 0; 0; 0
Max: 100; 100; 100

Practice/group level:
Age: 18-65
Measurement Year: 2016; 2017; 2018
Number of groups: 495; 581; 592
Mean: 86.14; 86.64; 85.99
Std Dev: 16.25; 16.08; 16.72
Min: 0; 0; 0
Max: 100; 100; 100

Age: >65
Measurement Year: 2016; 2017; 2018
Number of groups: 495; 581; 592

Mean: 100; 100; 100
Std Dev: 0; 0; 0
Min: 100; 100; 100
Max: 100; 100; 100

Race: White
Measurement Year: 2016; 2017; 2018
Number of groups: 495; 581; 592
Mean: 87.32; 87.74; 87.00
Std Dev: 19.39; 16.33; 17.47
Min: 0; 0; 0
Max: 100; 100; 100

Race: Black
Measurement Year: 2016; 2017; 2018
Number of groups: 495; 581; 592
Mean: 92.11; 89.35; 88.28
Std Dev: 17.59; 16.80; 18.26
Min: 0; 0; 0
Max: 100; 100; 100

Race: Asian
Measurement Year: 2016; 2017; 2018
Number of groups: 495; 581; 592
Mean: 88.58; 90.66; 91.22
Std Dev: 21.99; 20.50; 20.95
Min: 0; 0; 0
Max: 100; 100; 100

Race: Other
Measurement Year: 2016; 2017; 2018
Number of groups: 495; 581; 592
Mean: 87.76; 91.62; 89.75
Std Dev: 19.66; 18.40; 17.67
Min: 0; 0; 0
Max: 100; 100; 100

Race: Unknown
Measurement Year: 2016; 2017; 2018
Number of groups: 495; 581; 592
Mean: 88.87; 89.49; 88.80

Std Dev: 22.08; 16.72; 17.67
Min: 0; 0; 0
Max: 100; 100; 100

Ethnicity: Hispanic Latino
Measurement Year: 2016; 2017; 2018
Number of groups: 495; 581; 592
Mean: 87.70; 89.53; 89.64
Std Dev: 18.15; 20.41; 20.78
Min: 0; 0; 0
Max: 100; 100; 100

Ethnicity: Non-Hispanic Latino
Measurement Year: 2016; 2017; 2018
Number of groups: 495; 581; 592
Mean: 87.27; 88.03; 86.86
Std Dev: 18.31; 17.10; 18.86
Min: 0; 0; 0
Max: 100; 100; 100

Ethnicity: Unknown
Measurement Year: 2016; 2017; 2018
Number of groups: 495; 581; 592
Mean: 86.47; 88.97; 88.32
Std Dev: 16.64; 16.37; 16.85
Min: 0; 0; 0
Max: 100; 100; 100

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

This measure is not derived from patient report.

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

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

1a.3. SYSTEMATIC REVIEW(SR) OF THE EVIDENCE (for INTERMEDIATE OUTCOME, PROCESS, OR STRUCTURE PERFORMANCE MEASURES, INCLUDING THOSE THAT ARE INSTRUMENT-BASED) If the evidence is not based

on a systematic review go to section 1a.4) If you wish to include more than one systematic review, add additional tables.

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

☐ Clinical Practice Guideline recommendation (with evidence review)

X US Preventive Services Task Force Recommendation

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

X Other

The review of evidence provided during the initial endorsement is still current and relevant. This measure has been in the PQRS and MIPS program since 2013. We gave updated the evidence that was provided during the initial endorsement.

Source of Systematic Review: • **Title** • **Author** • **Date** • **Citation, including page number** • **URL**	• **Final Recommendation Statement *Colorectal Cancer: Screening*** • **USPSTF** • **June 21, 2016** • **JAMA. 2016;315(23):2564-2575. doi:10.1001/jama.2016.5989** • **https://www.uspreventiveservicestaskforce.org/Page/Document/RecommendationStatementFinal/colorectal-cancer-screening2** • **https://jamanetwork.com/journals/jama/fullarticle/2529486**
Quote the guideline or recommendation verbatim about the process, structure or intermediate outcome being measured. If	The USPSTF recommends screening for colorectal cancer starting at age 50 years and continuing until age 75 years.

not a guideline, summarize the conclusions from the SR.	
Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	Grade A The USPSTF recommends the service. There is high certainty that the net benefit is substantial. Practice Suggestion: Offer or provide this service.
Provide all other grades and definitions from the evidence grading system	A The USPSTF recommends the service. There is high certainty that the net benefit is substantial. Practice Suggestion: Offer or provide this service. 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. Practice Suggestion: Offer or provide this service. 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. Practice Suggestion: Offer or provide this service for selected patients depending on individual circumstances. 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. Practice Suggestion: Discourage the use of this service. 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. Practice Suggestion: Read the clinical considerations section of USPSTF Recommendation Statement. If the service is offered, patients should understand the uncertainty about the balance of benefits and harms.
Grade assigned to the **recommendation** with definition of the grade	Grade A The USPSTF recommends selectively offering or providing this service to individual patients based on professional judgment and patient preferences. There is at least moderate certainty that the net benefit is small. Offer or provide this service for selected patients depending on individual circumstances.
Provide all other grades and definitions from the recommendation grading system	High 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. Moderate The available evidence is sufficient to determine the effects of the preventive service on health outcomes, but confidence in the estimate is constrained by such factors as:

	The number, size, or quality of individual studies. Inconsistency of findings across individual studies. Limited generalizability of findings to routine primary care practice. Lack of coherence in the chain of evidence. As more information becomes available, the magnitude or direction of the observed effect could change, and this change may be large enough to alter the conclusion. Low The available evidence is insufficient to assess effects on health outcomes. Evidence is insufficient because of: The limited number or size of studies. Important flaws in study design or methods. Inconsistency of findings across individual studies. Gaps in the chain of evidence. Findings not generalizable to routine primary care practice. Lack of information on important health outcomes. More information may allow estimation of effects on health outcomes.
Body of evidence: • Quantity – how many studies? • Quality – what type of studies?	Published in June 2016: "The USPSTF commissioned a systematic evidence review to update its 2008 recommendation on screening for colorectal cancer. The review addressed the following: 1) the effectiveness of screening with colonoscopy, flexible sigmoidoscopy, CT colonography, gFOBT, FIT, FIT-DNA, and methylated SEPT9 DNA testing in reducing incidence of and mortality from colorectal cancer or all-cause mortality; 2) the harms of these screening tests; and 3) the test performance characteristics of these tests for detecting adenomatous polyps, advanced adenomas based on size, or both, as well as colorectal cancer. In contrast to the evidence review performed for the USPSTF in 2008, this review expanded its approach to additionally search for and consider 1) observational evidence about the benefits of screening tests when trial evidence does not exist and 2) comparative effectiveness of screening tests on cancer incidence and mortality. In addition, the USPSTF commissioned a report from the CISNET Colorectal Cancer Working Group to provide information from comparative modeling on optimal starting and stopping ages and screening intervals across the different available screening methods. Compared with the previous decision analysis performed for the USPSTF, this analysis used more narrowly defined ages at which to begin and end screening and screening intervals. It also included new screening methods (FIT-DNA, CT colonography, and flexible sigmoidoscopy combined with FIT), updated test characteristics, and age-specific risks of colonoscopy complications."
Estimates of benefit and consistency across studies	"The USPSTF found convincing evidence of benefit associated with colorectal cancer screening. The Hemoccult II test was the first colorectal cancer screening test to demonstrate reduction in disease-specific mortality in an RCT. Six trials showed that after 11 to 30 years of follow-up, screening with low-sensitivity gFOBT reduced the risk of colorectal cancer death by about 9% to 22% when performed biennially (about 9 to 16 fewer colorectal cancer deaths per 100,000

	person-years) and by about 32% when done annually.When considering the life-years gained compared with the burden and harms of screening (as assessed by the proxy measure of total number of lifetime colonoscopies), annual screening with high-sensitivity gFOBT was consistently dominated by annual FIT screening in the CISNET modeling."
What harms were identified?	"Screening with FIT-DNA and CT colonography each has several unique harms to consider. Screening with FIT-DNA is less specific than screening with FIT resulting in more false-positive results per screening test and an increased probability of harm from diagnostic colonoscopy. Further, a theoretical concern about FIT-DNA is whether its use might lead to more frequent and invasive follow-up testing in persons who are not at increased risk of colorectal cancer because of patient or clinician concerns about abnormal DNA results. Although modeling can be used to understand the estimated effects of the test's reduced specificity and increased false-positive rate, empirical evidence on appropriate follow-up of abnormal results is lacking, making it difficult to accurately understand the overall balance of benefits and harms of this screening test." "Screening with FIT-DNA and CT colonography each has several unique harms to consider. Screening with FIT-DNA is less specific than screening with FIT resulting in more false-positive results per screening test and an increased probability of harm from diagnostic colonoscopy. Further, a theoretical concern about FIT-DNA is whether its use might lead to more frequent and invasive follow-up testing in persons who are not at increased risk of colorectal cancer because of patient or clinician concerns about abnormal DNA results. Although modeling can be used to understand the estimated effects of the test's reduced specificity and increased false-positive rate, empirical evidence on appropriate follow-up of abnormal results is lacking, making it difficult to accurately understand the overall balance of benefits and harms of this screening test." "The direct harms of endoscopy have been somewhat better studied. Pooled estimates suggest there are about 4 (95% CI, 2 to 5) colonic perforations and about 8 (95% CI, 5 to 14) major intestinal bleeding episodes per 10,000 screening colonoscopies performed. Many of these events appear to be related to polypectomy, and the risk of experiencing an adverse event increases with age. The risk of bleeding or perforation seems to be greater if the colonoscopy is done as part of diagnostic follow-up of a positive finding on a screening test of a different method; for example, pooled data from flexible sigmoidoscopy trials found about 14 (95% CI, 9 to 26) colonic perforations and 24 (95% CI, 5 to 63) major bleeding episodes per 10,000 persons undergoing diagnostic colonoscopy. This compares to about 1 perforation and 2 major bleeding episodes per 10,000 flexible sigmoidoscopies performed for the purposes of cancer screening." "The harms from a single administration of a screening test must be considered in the context of how often the test will be repeated over a patient's lifetime. In the case of colorectal cancer screening, this means considering how many colonoscopies (the primary source of serious harms) will be required to follow up abnormal findings. The CISNET models suggest that the available strategies range from an estimated 1,714 to 4,049 total colonoscopies required per 1,000 persons screened over a lifetime; screening colonoscopy every 10 years generates the highest degree of associated burden or harm"
Identify any new studies conducted since the SR. Do the new studies	Neoplasia at 10-year follow-up screening colonoscopy in a private U.S. practice: comparison of yield to first-time examinations Rex, Douglas K. et al. Gastrointestinal Endoscopy, Volume 87, Issue 1, 254 - 259

change the conclusions from the SR?	Heisser Thomas, Peng Le, Weigl Korbinian, Hoffmeister Michael, Brenner Hermann. Outcomes at follow-up of negative colonoscopy in average risk population: systematic review and meta-analysis BMJ 2019; 367 :l6109 Yield of a second screening colonoscopy 10 years after an initial negative examination in average-risk individuals Ponugoti, Prasanna L. et al. Gastrointestinal Endoscopy, Volume 85, Issue 1, 221 - 224

Source of Systematic Review: • **Title** • **Author** • **Date** • **Citation, including page number** • **URL**	• **Colorectal Cancer Screening: Recommendations for Physicians and Patients from the U.S. Multi-Society Task Force on Colorectal Cancer** • **Rex, DK et al.** • **July 2017** • **Am J Gastroenterol. 2017 Jul;112(7):1016-1030. doi: 10.1038/ajg.2017.174. Epub 2017 Jun 6**
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.	We recommend colonoscopy every 10 years or annualFIT as first-tier options for screening the average-risk persons for colorectal neoplasia (strong recommendation; moderate quality evidence).
Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	Moderate-quality evidence
Provide all other grades and definitions from the evidence grading system	A: High quality Further research is very unlikely to change our confidence in the estimate of effect B: Moderate quality Further research is likely to have an important impact on our confidence in the estimate of effect and may change the estimate C: Low quality Further research is very likely to have an important impact on our confidence in the estimate of effect and is likely to change the estimate D: Very low quality Any estimate of effect is very uncertain
Grade assigned to the **recommendation** with definition of the grade	Strong recommendation (would be chosen by most informed patients)

Provide all other grades and definitions from the recommendation grading system	"Weak recommendations" are those where patient values and preferences might play a larger role than the quality of evidence. Within the document we preface strong recommendations with phrases such as "we recommend" and weak recommendations with "we suggest."
Body of evidence: • Quantity – how many studies? • Quality – what type of studies?	"Although no randomized trials of colonoscopy for screening have been completed, extensive evidence from adenoma cohorts, cohort studies on incidence and mortality, and case-control studies support the efficacy of colonoscopy in preventing incident CRC and cancer deaths. One cohort study and 3 case-control studies were performed in screening populations." "Furthermore, indirect evidence from randomized trials of fecal occult blood testing and sigmoidoscopy, as well as studies showing highly variable cancer protection provided by different colonoscopists, also supports a protective eff ect of colonoscopy against CRC. These fi ndings are consistent with the observed population trends in the United States."
Estimates of benefit and consistency across studies	No published randomized trials have directly compared and reported the relative effects of different tests on CRC incidence or mortality. Several trials are ongoing, but results are not yet available.
What harms were identified?	"There is currently insufficient evidence to recommend systematic screening in asymptomatic persons <50 years old who lack specific risk factors related to family history or Lynch syndrome. The yield of screening colonoscopy in this age group is low in available studies, and the biologic reasons for the increasing incidence of CRC in persons under age 50 years are uncertain. Additional study of the benefits and harms of screening in persons <50 years is warranted, perhaps particularly in persons with known colorectal risk factors such as cigarette smoking, diabetes mellitus, and obesity"
Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	Neoplasia at 10-year follow-up screening colonoscopy in a private U.S. practice: comparison of yield to first-time examinations Rex, Douglas K. et al. Gastrointestinal Endoscopy, Volume 87, Issue 1, 254 - 259 Heisser Thomas, Peng Le, Weigl Korbinian, Hoffmeister Michael, Brenner Hermann. Outcomes at follow-up of negative colonoscopy in average risk population: systematic review and meta-analysis BMJ 2019; 367 :l6109 Yield of a second screening colonoscopy 10 years after an initial negative examination in average-risk individuals Ponugoti,

	Prasanna L. et al. Gastrointestinal Endoscopy, Volume 85, Issue 1, 221 - 224

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

Based on Qaseem et al. Annals of Internal Medicine 2019:, which reviewed all guidelines related to colorectal cancer screening including the USMSTF 2017, USPSTF 2016: "No RCT data were available to determine the clinical benefits, including effects on CRC incidence or CRC-related and all-cause mortality. Indirect evidence from RCTs of flexible sigmoidoscopy, which allows direct visualization of the descending colon, suggests a CRC-specific mortality benefit. Modeling studies used in the USPSTF guideline also suggest such a benefit."

USPSTF 2016: Modeling data are included in the guidelines for screening colonoscopy every 10 years.

USMSTF 2017: While the number and type of study designs are not described by the guideline developers, the article did say, "Most of the information supporting the use of the other colorectal screening tests [including CSPY] is based on observational and inferential evidence. In this review, priority was placed on studies of asymptomatic average-risk or higher-risk populations that were followed by testing with colonoscopy in all or nearly all study participants as a validation measure."

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

USMSTF 2017: The guidance statements of the USMSTF were developed using the GRADE-based approach. Screening colonoscopy every 10 years is a tier 1 recommendation (strong recommendation; moderate quality of evidence).

From Qaseem et al Annals Int Med 2019:

"We include colonoscopy as an option for screening, as does the USPSTF, because indirect evidence (not from RCTs) suggests an association between reduced CRC mortality and colonoscopy compared with other options."

"Effectiveness of colonoscopy has not been evaluated in RCTs, but it is associated with the best sensitivity (67% to 94%) and specificity (96% to 98%) for adenomas measuring at least 10 mm and has been widely used for CRC screening on the basis of observational and modeling data. In addition, CRC mortality benefits associated with flexible sigmoidoscopy can be considered strong indirect evidence for colonoscopy benefits because both screening tests use direct visualization. Screening colonoscopy is currently recommended every 10 years (if results are normal)."

"Colonoscopy is an invasive procedure that requires bowel preparation and time spent attending an outpatient examination, and it is typically done using moderate sedation." Furthermore, "USPSTF-

estimated rate was similar at 4 perforations (CI, 2 to 5 perforations) in 10 000 procedures. Follow-up colonoscopy after positive findings on flexible sigmoidoscopy screening resulted in 14 perforations (CI, 9 to 26 perforations) per 10 000 procedures and 34 major bleeding events (CI, 5 to 63 events) per 10 000 procedures." Also, the risk for major bleeding for which hospitalization was required was, "estimated as 8.21 events (CI, 4.98 to 13.51 events) per 10 000 procedures by the USPSTF." Finally, " The USPSTF notes that cardiopulmonary adverse events may occur with colonoscopy if sedation is used but that the frequency is unknown."

Rex et al. USMSTF 2017:

"Although no randomized trials of colonoscopy for screening have been completed, extensive evidence from adenoma cohorts (Winawer 1993, Zauber 2012), cohort studies on incidence and mortality (Singh 2010, Kahi 2009), and case-control studies (57–64) support the efficacy of colonoscopy in preventing incident CRC and cancer deaths. One cohort study (56) and 3 case-control studies (58,59,64) were performed in screening populations. Reductions in incidence and mortality are approximately 80% in the distal colon and 40 to 60% in the proximal colon, at least in the United States and Germany (57,59,61,62,64). Furthermore, indirect evidence from randomized trials of fecal occult blood testing (65) and sigmoidoscopy (66), as well as studies showing highly variable cancer protection provided by different colonoscopists (67,68), also supports a protective effect of colonoscopy against CRC. These findings are consistent with the observed population trends in the United States (11,12).

USMSTF 2008: Again, while the magnitude and direction across studies was not described, the guideline developers did summarize other studies as follows: "The evaluation of incidence rates of CRC in adenoma cohorts after baseline CSPY and polypectomy is another form of evidence commonly cited to support CSPY for CRC screening. In the National Polyp Study, the incidence of CRC after clearing CSPY was reduced by 76% to 90% compared with 3 nonconcurrent reference populations. In an Italian adenoma cohort study with removal of at least one adenoma ≥5 mm, there was an 80% reduction in CRC incidence compared with expected incidence in a reference population. However, not all studies have shown the same level of protection. Combined data from 3 US chemoprevention trials showed incidence rates of CRC after clearing CSPY approximately 4 times that seen in the National Polyp Study, with no reduction in CRC incidence compared with data from the Surveillance Epidemiology and End Results (SEER) database in the United States, and 2 US dietary intervention trials also showed higher rates of incident CRC after clearing CSPY than were observed in the National Polyp Study. These differences may reflect exclusion of patients with sessile adenomas >3 cm in the National Polyp Study, more effective baseline clearing (13% of patients in the National Polyp Study had 2 or more baseline CSPY to complete clearing), or unmeasured differences in the average quality of CSPY between the studies. Overall, the data support the conclusion that CSPY with clearing of neoplasms by polypectomy has a significant impact on CRC incidence and thus, by extension, mortality. The magnitude of the protective impact is uncertain; it is not absolute, nor are apparent failures well understood. In a study of 35,000 symptomatic patients in Manitoba who had undergone a negative CSPY and who then were followed for 10 years, the investigators observed significant reductions in CRC incidence over time, but the incidence reductions were less than 50% for each of the first 5 years and no more than 72% by 10 years. These findings suggest detection failures during the initial, apparently normal, CSPY."

USMSTF 2008: The guideline developers have identified the following harms that have been studied, which they deem minimal in comparison to the benefits: "Controlled studies have shown the CSPY miss rate for large adenomas (≥10 mm) to be 6% to 12%. The reported CSPY miss rate for cancer is about 5%. CSPY can result in significant harm, most often associated with polypectomy, and the most common serious complication is post polypectomy bleeding. The risk of post polypectomy bleeding is increased with large polyp size and proximal colon location; however, small polyp bleeds are more numerous than large polyp bleeds because small polyps are so numerous. Another significant risk associated with CSPY is perforation. Perforation increases with

increasing age and the presence of diverticular disease and was recently estimated to occur in 1 in 500 of a Medicare population and approximately 1 in 1000 screened patients overall.123 Because of the age effect, perforation rates measured in the Medicare population may overestimate the overall risk of perforation in CSPY; however, a large study in the Northern California Kaiser Permanente population also identified a perforation rate of 1 in 1000. In addition, cardiopulmonary complications such as cardiac arrhythmias, hypotension, and oxygen desaturation may occur, although these events rarely result in hospitalization. Cardiopulmonary complications represent about one half of all adverse events that occur during CSPY and usually are related to sedation. Thus, while screening CSPY has established benefits with regard to the detection of adenomas and cancer, complications related to CSPY are a significant public health challenge." However, despite these risks of harm, "A principal benefit of CSPY is that it allows for a full structural examination of the colon and rectum in a single session and for the detection of colorectal polyps and cancers accompanied by biopsy or polypectomy. All other forms of screening, if positive, require CSPY as a second procedure. Patient surveys indicate that patients willing to undergo invasive testing tend to choose CSPY as their preferred test. In addition to being a complete examination of the colon, individuals may also regard sedation during the procedure as an advantage. Patients in the same practice who had undergone unsedated FSIG screening were more than twice as likely to say that they would not return for additional screening compared with those who had undergone CSPY with sedation."

Rex, et al, 2009: The magnitude and direction across studies was not described, but the guideline developers summarized the benefits of a number of studies as follows: "The evidence that colonoscopy prevents incident CRCs and reduces the consequent mortality from CRC is indirect but substantial. No prospective randomized controlled trial, comparing colonoscopy with no screening, has been carried out. However in a randomized controlled trial, involving only 800 patients, in which flexible sigmoidoscopy with colonoscopy carried out for any polyp detected was compared with no screening, the screening strategy resulted in an 80 % reduction in the incidence of CRC. In addition, at the University of Minnesota, a randomized controlled trial was carried out comparing annual vs. biennial fecal occult blood testing with rehydration with no screening. Screening resulted in a 20% incidence reduction in CRC, which appeared to have resulted from detection of large adenomas by fecal occult blood testing and subsequent colonoscopy and polypectomy. Cohort studies involving patients, who have undergone colonoscopy and polypectomy with apparent clearance of colonic neoplasia, have shown a 76 – 90% reduction in the incidence of CRC in comparison with reference populations. Case – control studies of colonoscopy showed a 50% reduction in mortality from CRC in a US Veterans Administration population, and there was an 80% reduction in the CRC incidence in the German population . Population-based studies in the United States have associated increases in the use of colonoscopy with earlier and more favorable stages in CRC presentation , and with reductions in the incidence of CRC. Additional evidence for a benefit from colonoscopy screening is extrapolated from case – control studies of sigmoidoscopy, which have shown mortality and incidence reductions of distal CRC of 60 and 80%, respectively, in screening populations."

Colorectal cancer is the 2nd leading cause of cancer-related death in the United States. Inappropriate interval recommendations can result in overuse of resources and can lead to significant patient harm. Performing colonoscopy too often not only increases patients' exposure to procedural harm, but also drains resources that could be more effectively used to adequately screen those in need (Lieberman et al, 2009).

A recent community based multi-organ cancer screening study in 3627 patients noted that 49 % of low risk patients with adequate negative colonoscopic examinations underwent follow-up surveillance procedures within 7 years (median 3.1 yrs) of their first study, and 35% of low risk patients with two negative exams underwent a third study at a median of 3.3 years after the prior study, despite guidelines for repeat examination at 10 years (Schoen, 2010). Variations in the recommended time interval between colonoscopies also exist for patients with normal colonoscopy findings. In a 2006 study of 1282 colonoscopy reports, recommendations were consistent with current guidelines in only 36.7% of cases. (Krist et al, 2007).

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

The evidence review was previously submitted by PCPI in 2012 during the GI/GU project and continues to be relevant. The measure has been maintained in the PQRS and MIPS program annually and any changes in evidence are continually reviewed during the measure maintenance cycles. There have been no significant changes in evidence since the initial endorsement in 2012.

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

At present, CSPY (colonoscopy) every 10 years is an acceptable option for CRC screening in average-risk adults beginning at age 50 years. (USMSTF 2017)

One of the tier 1 approaches for CRC prevention is colonoscopy every 10 years, beginning at age 50. (Strong recommendation; moderate quality of evidence) (Rex et al, 2017)

Zauber, et al. Evaluating test strategies for colorectal cancer screening; a decision analysis for the US preventive services task force. Ann Int Med Vol 149, 2008.

Lieberman, DA, Faigel, DO, Logan, J, Mattek, N, Holub, J, Eisen, G, Morris, C, Smith, R, Nadel, M. Assessment of the Quality of Colonoscopy Reports: Results from a multi-center consortium. Gastrointest Endosc Vol 69, 2009.

Schoen R, Pinsky PF, Weissfeld JL, et al. Utilization of Surveillance Colonoscopy in Community Practice. Gastroenterology Vol 138, 2010.

Krist, AH, jones, RM, Woolf, SH et al. Timing of Repeat Colonoscopy: Disparity Between Guidelines and Endoscopists' Recommendation. American Journal of Preventive Medicine. 2007.

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)*

<u>If a COMPOSITE</u> (e.g., combination of component measure scores, all-or-none, any-or-none), SKIP this question and answer the composite questions.

Guideline recommendations support screening colonoscopy at 10 year intervals, for average risk patients. Non-adherence to guideline recommendations increases patients to unnecessary risk via procedural harms and complications. Colonoscopy screening at more frequent intervals also contributes to increased costs to patients and insurers.

In the average-risk population, colonoscopy screening is recommended in all current guidelines at 10-year intervals. Inappropriate interval recommendations can result in overuse of resources and can lead to significant patient harm. Performing colonoscopy too often not only increases patients' exposure to procedural harm, but also drains resources that could be more effectively used to adequately screen those in

need (Lieberman et al, 2008). The most common serious complication of colonoscopy is post-polypectomy bleeding (Levin et al, 2008).

Variations in the recommended time interval between colonoscopies exist for patients with normal colonoscopy findings. In a 2006 study of 1282 colonoscopy reports, recommendations were consistent with contemporaneous guidelines in only 39.2% of cases and with current guidelines in 36.7% of cases. Further, the adjusted mean number of years in which repeat colonoscopy was recommended was 7.8 years following normal colonoscopy (Krist et al, 2007)

1b.2. Provide performance scores on the measure as specified (<u>current and over time</u>) at the specified level of analysis. *(<u>This is required for maintenance of endorsement</u>. 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.*

A recent community based multi-organ cancer screening study in 3627 patients noted that 49 % of low risk patients with adequate negative colonoscopic examinations underwent follow-up surveillance procedures within 7 years (median 3.1 yrs) of their first study, and 35% of low risk patients with two negative exams underwent a third study at a median of 3.3 years after the prior study, despite guidelines for repeat examination at 10 years (Schoen, 2010). Variations in the recommended time interval between colonoscopies also exist for patients with normal colonoscopy findings. In a 2006 study of 1282 colonoscopy reports, recommendations were consistent with current guidelines in only 36.7% of cases. (Krist et al, 2007).

The data provided to demonstrate a continued opportunity for improvement is from the GI Quality Improvement Consortium, Ltd (GIQuIC), a non-profit collaboration of the American College of Gastroenterology (ACG) and the American Society for Gastrointestinal Endoscopy (ASGE), which established the GIQuIC clinical benchmarking registry, an approved Qualified Clinical Data Registry (QCDR). Performance from 2016-2018 is provided at the individual physician and practice/group level with the mean, standard deviation, minimum and maximum performance as well as the interquartile range (IQR).

Individual physician

Measurement Year: 2016; 2017; 2018

Number of physicians: 3,136; 3,618; 3,747

Mean: 85.12; 85.63; 85.43

Std Dev: 23.71; 23.32; 23.21

Min: 0; 0; 0

Max: 100; 100; 100

IQR: 15.36; 14.99; 15.67

Practice/group

Measurement Year: 2016; 2017; 2018

Number of physicians: 495; 581; 592

Mean: 87.79; 88.29; 87.71

Std Dev: 15.18; 15; 15.54

Min: 0; 0; 0

Max: 100; 100; 100

IQR: 11.72; 11.43; 12.20

The data provided to demonstrate a continued opportunity for improvement is from GI Quality Improvement Consortium, Ltd (GIQuIC), a non-profit collaboration of the American College of Gastroenterology (ACG) and the American Society for Gastrointestinal Endoscopy (ASGE), which established the GIQuIC clinical

benchmarking registry, an approved Qualified Clinical Data Registry (QCDR). Overall performance on the measure from 2016-2018 stratified by age and race/ethnicity is provided with the mean, standard deviation, minimum and maximum performance at the individual physician and practice/group levels.

Individual Physician Level:

Age: 18-65

Measurement Year: 2016; 2017; 2018

Number of physicians: 3,142; 3,634; 3,763

Mean: 83.71; 834.19; 83.96

Std Dev: 24.37; 24.03; 23.97

Min: 0; 0; 0

Max: 100; 100; 100

Age: >65

Measurement Year: 2016; 2017; 2018

Number of physicians: 3,142; 3,634; 3,763

Mean: 100; 100; 100

Std Dev: 0; 0; 0

Min: 100; 100; 100

Max: 100; 100; 100

Race: White

Measurement Year: 2016; 2017; 2018

Number of physicians: 3,142; 3,634; 3,763

Mean: 85.08; 85.81; 85.48

Std Dev: 24.62; 23.61; 23.71

Min: 0; 0; 0

Max: 100; 100; 100

Race: Black

Measurement Year: 2016; 2017; 2018

Number of physicians: 3,142; 3,634; 3,763

Mean: 86.10; 87.26; 87.19

Std Dev: 27.01; 25.79; 25.57

Min: 0; 0; 0

Max: 100; 100; 100

Race: Asian

Measurement Year: 2016; 2017; 2018

Number of physicians: 3,142; 3,634; 3,763

Mean: 90.79; 89.88; 91.00

Std Dev: 24.96; 26.06; 28.37

Min: 0; 0; 0

Max: 100; 100; 100

Race: Other

Measurement Year: 2016; 2017; 2018

Number of physicians: 3,142; 3,634; 3,763

Mean: 86.26; 88.56; 87.33

Std Dev: 28.95; 26.84; 28.37

Min: 0; 0; 0

Max: 100; 100; 100

Race: Unknown

Measurement Year: 2016; 2017; 2018

Number of physicians: 3,142; 3,634; 3,763

Mean: 86.43; 87.32; 87.08

Std Dev: 25.54; 24.65; 24.89

Min: 0; 0; 0

Max: 100; 100; 100

Ethnicity: Hispanic Latino

Measurement Year: 2016; 2017; 2018

Number of physicians: 3,142; 3,634; 3,763

Mean: 89.19; 88.61; 88.95

Std Dev: 25.58; 26.90; 26.29

Min: 0; 0; 0

Max: 100; 100; 100

Ethnicity: Non-Hispanic Latino

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.

Schoen R, Pinsky PF, Weissfeld JL, et al. Utilization of Surveillance Colonoscopy in Community Practice. Gastroenterology Vol 138, 2010.

Krist, AH, jones, RM, Woolf, SH et al. Timing of Repeat Colonoscopy: Disparity Between Guidelines and Endoscopists' Recommendation. American Journal of Preventive Medicine. 2007.

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.*

After a search of the medical literature, we are not aware of any publications/evidence outlining disparities in this area.

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

N/A

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.

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):*

Gastrointestinal (GI)

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

Primary Prevention, Safety : Overuse, Screening

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

Elderly

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.)*

www.gastro.org

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

This is not an eMeasure **Attachment:**

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

No data dictionary **Attachment:**

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

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

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

Not an instrument-based measure

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

No

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

Age band of 50 to 75 years of age was created to conform with the USPTF recommendations during the 2016 maintenance cycle

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 had a recommended follow-up interval of at least 10 years for repeat colonoscopy documented in their colonoscopy report

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).

Patients will be counted in the numerator if it is documented in the final colonoscopy report that the appropriate follow-up interval for the next colonoscopy is at least 10 years from the date of the current colonoscopy (ie, the colonoscopy performed during the measurement period).

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

All patients aged 50 years to 75 years and receiving screening a screening colonoscopy without biopsy or polypectomy

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).

All patients aged 50 to 75 years of age receiving a screening colonoscopy without biopsy or polypectomy during the measurement period.

ICD-10-CM: Z12.11

AND

Patient encounter during the reporting period (CPT or HCPCS): 44388, 45378, G0121

WITHOUT

CPT Category I Modifiers: 52, 53, 73, 74

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

Documentation of medical reason(s) for not recommending at least a 10 year follow-up interval (eg, inadequate prep,familial or personal history of colonic polyps, patient had no adenoma and age is >= 66 years old, or life expectancy < 10 years, other medical reasons)

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.)

The measure exception categories are not uniformly relevant across all measures; for each measure, there must be a clear rationale to permit an exception for a medical, patient, or system reason. Examples are provided in the measure exception language of instances that may constitute an exception and are intended to serve as a guide to clinicians. For measure 0658, exceptions may include medical reason(s) (eg, inadequate prep, other medical reasons) for not recommending at least a 10 year follow-up interval. Examples of exceptions are included in the measure language.

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

We encourage the results of this measure to be stratified by race, ethnicity, gender, and primary language.

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

No risk adjustment or risk stratification

If other:

S.12. Type of score:

Rate/proportion

If other:

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

Better quality = Higher score

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

To calculate performance rates:

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

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

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

4) From the patients who did not meet the numerator criteria, determine if the physician has documented that the patient meets any criteria for denominator exception when exceptions have been specified [for this measure: medical reason(s) (eg, inadequate prep, familial or personal history of colonic polyps, patient had no adenoma and age is >= 66 years old, life expectancy < 10 years, other medical reasons)]. If the patient meets any exception criteria, they should be removed from the denominator for performance calculation. --Although the exception cases are removed from the denominator population for the performance calculation, the number of patients with valid exceptions should be calculated and reported along with performance rates to track variations in care and highlight possible areas of focus for QI.

If the patient does not meet the numerator and a valid exception is not present, this case represents performance not met.

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.

Not applicable. The measure does not require sampling or a survey.

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

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

N/A

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

If other, please describe in S.18.

Claims, Electronic Health Data, Electronic Health Records, Other, Registry Data

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

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

Not applicable.

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

No data collection instrument provided

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

Clinician : Individual

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

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

NQF_658_-_Testing_Attachment_FINAL-636426432396770942.doc,AGA_0658_Testing_Attachment_8-1-18.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)

NATIONAL QUALITY FORUM—Measure Testing (subcriteria 2a2, 2b1-2b6)

Measure Number (*if previously endorsed*)**:** 0658
Measure Title: Appropriate Follow-Up Interval for Normal Colonoscopy in Average Risk Patients
Date of Submission: 8/1/2018

Type of Measure:

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

Instructions

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

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

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

2b1. Validity testing [11] demonstrates that the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality. For instrument-based measures (including PRO-PMs) and composite performance measures, validity should be demonstrated for the computed performance score.

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

AND

If patient preference (e.g., informed 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: Click here to describe

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

Data 1

The data source is the GIQuIC (GI Quality Improvement Consortium, Ltd.) registry, a procedure focused benchmarking registry using established quality indicators.

Data was collected electronically via endowriter, an automated endoscopy record system (not an EHR/EMR) or manually via a web portal. Data can be reported to PQRS. Additionally, registry participants use the data for their unit quality improvement programs and can report the data to programs such as ASGE's Endoscopy Unit Recognition Program.
http://giquic.gi.org/

Data 2

The data source is the AGA Digestive Health Outcomes Registry, a procedure-focused benchmarking registry using established quality indicators. The data are collected via EMR as well as web-portal data entry. The EMR data are sourced through a certified data transmission and validation process. Data can be reported to PQRS. www.agaregistry.org

Data 1

The data source is registry data from the Physician Quality Reporting System (PQRS), provided by the Centers for Medicare & Medicaid Services (CMS).

Data 2

The data source is the GI Quality Improvement Consortium, Ltd (GIQuIC), a non-profit educational and scientific organization established by the American College of Gastroenterology (ACG) and the American Society for Gastrointestinal Endoscopy (ASGE) that is an approved Qualified Clinical Data Registry (QCDR).

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

Data 1

The data are for the time period July 2010-October 2012, and cover the entire United States.

Data 2

The data are for the time period January 2011 to December 2011, and cover the entire United States.

Data 1

The data are for the time period January 2016 through December 2016 and cover the entire United States. Given the required conversion to ICD-10 in late 2015, the testing was completed on the ICD-10 specified measure.

Data 2

The data are for the time period January 2016 through December 2016 and cover the entire United States. Given the required conversion to ICD-10 in late 2015, the testing was completed on the ICD-10 specified measure.

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

Measure Specified to Measure Performance of: (***must be consistent with levels entered in item S.20***)	**Measure Tested at Level of:**
☒ individual clinician	☒ individual clinician
☐ group/practice	☐ group/practice
☐ hospital/facility/agency	☐ hospital/facility/agency
☐ health plan	☐ health plan
☐ other: Click here to describe	☐ other: Click here to describe

1.5. How many and which measured entities were included in the testing and analysis (by level of analysis and data source)? (*identify the number and descriptive characteristics of measured entities included in the analysis (e.g., size, location, type); if a sample was used, describe how entities were selected for inclusion in the sample*)

For this measure, the minimum number required to be included is 10 events. Given the structure of the PQRS program, a physician may choose to submit or not submit to PQRS on any given claim. Since these data contain results on a large number of physicians, limiting the reliability analysis to only those physicians who are participating in the program will eliminate the bias introduced by the inclusion of from physicians who are in the data but are not submitting claims to PQRS.

Data 1
An additional use of the GIQuIC registry would be for participants to use the data for completing their Self-Directed Practice Improvement Module as part of their recertification with ABIM. Since we are limiting the analysis to only those with 10 or more events due to the structure of PQRS reporting, to maintain consistency, we are also limiting to physicians who have 10 or more events for the purpose of recertification with ABIM.

177 physicians had all the required data elements and met the minimum number of quality reporting events (10) for inclusion in the reliability analysis. The average number of quality reporting events for physicians included is 81.16 for a total of 14,366 events. The range of quality reporting events for physicians included is from 587 to 10.

97% of the physicians were associated with ambulatory endoscopy centers, 2 % were at hospitals, and 1 % was with an office based practice. The average number of physicians per site was 13.6 with a range of 1 to 27 physicians per site. The centers were located in 13 different states across the US.

Data 2
20 physicians had all the required data elements and met the minimum number of quality reporting events (10) for inclusion in the reliability analysis. The average number of quality reporting events for physicians included is 95.55 for a total of 1,911 events. The range of

quality reporting events for physicians included is from 389 to 12.

Data 1

We received data from 458 physicians reporting on this measure through the registry option for CMS's PQRS in 2016. Of those, 237 physicians had all the required data elements and met the minimum number of quality reporting events (10) for a total of 5,445 quality events with an average number of 23 quality reporting events per physician. The range of quality reporting events for the 237 physicians included is from 10 to 197. The average number of quality reporting events for the remaining 48 percent of physicians that aren't included is 4.

Data 2

We received data from 3,030 physicians through the GIQuIC in 2016. Of those, 2,666 physicians had all the required data elements and met the minimum number of quality reporting events (10) for a total of 229,209 quality events with an average number of 86 quality reporting events per physician. The range of quality reporting events for 2,666 physicians included is from 10 to 805. The average number of quality reporting events for the remaining 12 percent of physicians that aren't included is 5.

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*)

Data 1

There were 14,366 patients included in this testing and analysis. These were the patients that were associated with physicians who had 10 or more patients eligible for this measure.

The average age was 58.9 with a range from 50 to 93 years old. 61.5% of the sample was female, 38.5% male. Racial breakout was as follows:

Race	Percentage of Total	Percentage with Known Race
African American	8.47%	10.61%
Asian Pacific	1.60%	2.01%
Hispanic	3.57%	4.47%
White, Non-Hispanic	66.20%	82.91%
Unknown	20.16%	-

Data 2

There were 1,911 patients included in this testing and analysis. These were the patients that were associated with physicians who had 10 or more patients eligible for this measure.

Data 1

There were 5,445 patients included in this reliability testing and analysis. These were the patients that were associated with physicians who had 10 or more patients eligible for this measure.

Data 2

There were 229,209 patients included in this reliability testing and analysis. These were the patients that were associated with physicians who had 10 or more patients eligible for this measure.

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.

The same data sample from each registry was used for the respective reliability testing, performance testing, and exceptions analysis.

Data 1

The same data sample from CMS Registry reporting was used for the respective reliability testing, performance testing, and exceptions analysis.

Data 2

The same data sample from the GIQuIC registry was used for the respective reliability testing, performance testing, and exceptions analysis.

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.

Data 1

Patient-level socio-demographic (SDS) variables were not captured as part of the testing.

Data 2

Patient-level socio-demographic (SDS) variables were not captured as part of the testing.

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*)

Reliability of the computed measure score was measured as 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 physician performance. Reliability at the level of the specific physician is given by:

Reliability = Variance (physician-to-physician) / [Variance (physician-to-physician) + Variance (physician-specific-error]
Reliability is the ratio of the physician-to-physician variance divided by the sum of the physician-to-physician variance plus the error variance specific to a physician. 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 physician performance.

Reliability testing was performed by using a beta-binomial model. The beta-binomial model assumes the physician performance score is a binomial random variable conditional on the physician'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.

Reliability is estimated five different points: at the minimum number of quality reporting events for the measure; at the mean number of quality reporting events per physician; and at the 25th, 50th and 75th percentiles of the number of quality reporting events.

Reliability of the computed measure score was measured as 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 physician performance and the noise is the total variability in measured performance.

Reliability at the level of the specific physician is given by:

Reliability = Variance (physician-to-physician) / [Variance (physician-to-physician) + Variance (physician-specific-error]

Reliability is the ratio of the physician-to-physician variance divided by the sum of the physician-to-physician variance plus the error variance specific to a physician. 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 physician performance.

Reliability testing was performed by using a beta-binomial model. The beta-binomial model assumes the physician performance score is a binomial random variable conditional on the physician'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.

Reliability is evaluated by averaging over physician specific reliabilities for all providers that meet the minimum number of quality reporting events for the measure. Each provider must have at least 10 eligible reporting events to be included in this calculation.

A reliability equal to zero implies that all the variability in a measure is attributable to measurement error. A reliability equal to one implies that all the variability is attributable to real differences in physician performance. A reliability of 0.70 – 0.80 is generally considered the acceptable threshold for reliability, 0.80 – 0.90 is considered high reliability, and 0.90 – 1.0 is considered very high. [1]

1. Adams JL, Mehrotra A, McGlynn EA, Estimating Reliability and Misclassification in Physician Profiling, Santa Monica, CA: RAND Corporation, 2010. www.rand.org/pubs/technical_reports/TR863. (Accessed on February 24, 2012.)

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*)

Data 1

For this measure, the reliability at the minimum level of quality reporting events (10) was 0.79. The average number of quality reporting events for physicians included is 81.16. The reliability at the average number of quality reporting events was 0.97

Description	Number of events	Reliability
Average	81	0.969
Minimum	10	0.797
75th percentile	98	0.975
50th percentile	53	0.954
25th percentile	21	0.892

Data 2

For this measure, the reliability at the minimum level of quality reporting events (10) was 0.86. The average number of quality reporting events for physicians included is 95.55. The reliability at the average number of quality reporting events was 0.98

Description	Number of events	Reliability
Average	96	0.979
Minimum	10	0.855
75th percentile	135	0.983
50th percentile	28	0.969
25th percentile	18	0.925

Data 1

The reliability above the minimum level of quality reporting events was 0.90.

Data 2

The reliability above the minimum level of quality reporting events was 0.94.

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?*)

Data 1
This measure has moderate reliability when evaluated at the minimum level of quality reporting events and high reliability at the median number of events (50th percentile), and at average and greater number of quality events. This suggests that for physicians with an average or greater number of events the measure has high reliability.

Data 2
This measure has high reliability when evaluated at the minimum level of quality reporting events and high reliability at the median number of events (50th percentile), and at average and greater number of quality events. This suggests that for physicians with an average or greater number of events the measure has high reliability.

Data analyses were conducted by using SAS/STAT software, version 8.2 (SAS Institute, Cary, North Carolina).

Data 1
This measure has very high reliability when evaluated above the minimum level of quality reporting events.

Data 2
This measure has very high reliability when evaluated above the minimum level of quality reporting events.

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)*

All PCPI performance measures are assessed for content validity by a panel of expert work group members during the development process. Additional input on the content validity of draft measures is obtained through a 30-day public comment period and by also soliciting comments from a panel of consumer, purchaser, and patient representatives convened by the PCPI specifically for this purpose. All comments received are reviewed by the expert work group and the measures adjusted as needed. Other external review groups (eg, focus groups) may be convened if there are any remaining concerns related to the content validity of the measures.

An expert panel was used to systematically assess face validity of the measure. The panel was asked to rate their agreement with the following statement:

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

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

The expert panel included 21 members from the following specialty areas: gastroenterology, colon and rectal surgery, general surgery, health plans, internal medicine, pathology, family medicine, infectious diseases and medical informatics.

John Allen, MD, MBA, AGAF (Gastroenterology), Minneapolis, MN
Doug Faigel, MD (Gastroenterology), Scottsdale, AZ
Nancy Baxter, MD, PhD, FACRS, FACS (Colon and Rectal Surgery) Arlington Heights, IL
Stephen Bickston, MD, AGAF (Gastroenterology)
Joel V. Brill, MD, AGAF, FASGE, FACG, CHCQM (Gastroenterology), Phoenix, AZ
Kirk Brandon, MBA (Business Administration/Coding)
Jason A. Dominitz, MD, MHS, AGAF (Gastroenterology) VA Puget Sound Health Care System, Seattle, WA
Ira L. Flax, MD, FACG (Gastroenterology) American College of Gastroenterology, Houston, TX
Karen E. Hall, MD, PhD (Geriatrics) University of Michigan HS, Ann Arbor, MI
Robert Haskey, MD, FACS (General Surgery, Health Plan representative)
Brian C. Jacobson, MD, MPH (Gastroenterology) ASGE, Needham, MA
David Lieberman, MD (Gastroenterology)
Klaus Mergener, MD, PhD, CPE, FACP, FACG, FASGE, FACPE (Gastroenterology) Tacoma, WA
Bret Petersen, MD, FASGE (Gastroenterology), Rochester, MN
Irving M. Pike, MD, FACG (Gastroenterology), Virginia Beach, VA
Bart Pope, MD (Family Medicine)
Harry Sarles, MD, FACG (Gastroenterology)
Kay Schwebke, MD, MPH (Internal Medicine, Infectious Diseases & Medical Informatics)
Optum Insight, Eden Prairie, MN
Tom Lynn, MD (Medical Informatics, Methodology)
Emily E. Volk, MD, FCAP (Pathology) San Antonio, TX
Michael Weinstein, MD (Gastroenterology) Chevy Chase, MD

To satisfy NQF's ICD-10 Conversion Requirements, we are providing the information below:

- NQF ICD-10-CM Requirement 1: Statement of intent related to ICD-10 CM
  Goal was to convert this measure to a new code set, fully consistent with the original intent of the measure.

- NQF ICD-10-CM Requirement 2: Coding Table
  See attachment in S.2b

- NQF ICD-10-CM Requirement 3: Description of the process used to identify ICD-10 codes
  The PCPI uses the General Equivalence Mappings (GEMs) as a first step in the identification of ICD-10 codes. We then review the ICD-10 codes to confirm their inclusion in the measure is consistent with the measure intent, making additions or deletions as needed. We have an RHIA-credentialed professional on our staff who reviews all ICD-10 coding. For measures included in CMS' Quality Payment Program (QPP), the ICD-10 codes have also been reviewed and vetted by the CMS contractor. Comments received from stakeholders related to ICD-10 coding are first reviewed internally. Depending on the nature of the comment received, we also engage clinical experts to advise us as to whether a change to the specifications is warranted.

Data 1

Colorectal Cancer Screening (PQRS #113) was chosen as a suitable candidate for correlation analysis due to the similarities in patient population and domain. We hypothesize that there exists a positive association between patients receiving a screening colonoscopy (PQRS #113) and those who had documentation of appropriate recommended follow-up interval of at least 10 years for repeat colonoscopy (NQF #0658). Providers included in the analysis met the minimum number of quality reporting events (10) and were cleaned in the same process as the PQRS dataset.

Datasets were reviewed to identify shared providers based on randomly generated identifiers in place of NPI and TIN identifiers. Correlation analysis was then performed to evaluate the association between performance scores of these shared providers. We use the following guidance to describe correlation[1]:

Correlation	Interpretation
> 0.40	Strong
0.20 - 0.40	Moderate
< 0.20	Weak

1. Shortell T. An Introduction to Data Analysis & Presentation. Sociology 712. http://www.shortell.org/book/chap18.html. Accessed July 13, 2018.

Data 2

Correlation analysis could not be performed on this data set.

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

The aforementioned expert panel was used to systematically assess face validity of the measure. They were asked to rate their agreement with the following statement:
"The scores obtained from the measure as specified will accurately differentiate quality across providers."

Scale 1-5, where 1=Strongly Disagree; 3=Neither Disagree nor Agree; 5=Strongly Agree
The results of the expert panel rating of the validity statement for Measure 658 were as follows: N = 14; Mean rating = 4.36 and 92.86% of respondents either agree or strongly agree that this measure can accurately distinguish good and poor quality.
Frequency Distribution of Ratings
1 - 1 (Strongly Disagree)
2 - 0
3 – 0 (Neither Agree nor Disagree)
4 - 5
5 - 8 (Strongly Agree)

Data 1

Appropriate Follow-Up Interval for Normal Colonoscopy in Average Risk Patients was positively correlated with Colorectal Cancer Screening (PQRS #113).

PQRS #113

Coefficient of correlation = 0.20

P-value = 0.007

Data 2

N/A

2b1.4. What is your interpretation of the results in terms of demonstrating validity? (i.*e., what do the results mean and what are the norms for the test conducted?*)
The results of the expert panel rating of the validity statement for Measure 658 were as follows: N = 14; Mean rating = 4.36 and 92.86% of respondents either agree or strongly agree that this measure can accurately distinguish good and poor quality.
These results demonstrate that Measure 658 has high face validity.

Data 1

Appropriate Follow-Up Interval for Normal Colonoscopy in Average Risk Patients has a moderate positive correlation with another evidence-based process of care measure. The correlation is statistically significant at the 95% confidence level and demonstrates the criterion validity of the measure.

Data 2

N/A

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*)

Exceptions were determined based on reported characteristics of the endoscopy. Some of the possible reasons for a denominator exception could be: inadequate bowel prep; incomplete colon examination; above average patient risk; complications arising during colonoscopy.

The examples are congruent with guidance from the ASGE in their 2006 guidelines for colorectal cancer screening and surveillance which indicate that "the completeness of the examination and the quality of the preparation should be taken into account for the timing of subsequent examinations."

Exceptions were determined based on reported characteristics of the endoscopy. Some of the possible reasons for a denominator exception could be: inadequate bowel prep; incomplete colon examination; above average patient risk; complications arising during colonoscopy.
The examples are congruent with guidance from the ASGE in their 2006 guidelines for colorectal cancer screening and surveillance which indicate that "the completeness of the examination and the quality of the preparation should be taken into account for the timing of subsequent examinations."

Exceptions were analyzed for frequency and variability across providers.

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*)

Data 1
For the 177 physicians that had all the required data elements and met the minimum number of quality reporting events (10) for inclusion in the reliability analysis, there were a total of 17,640 quality reporting events. 3,274 of the events were considered exceptions for an exception rate of 18%. The average number of exceptions for 177 physicians included is 18.5. The range of exception rates for physicians included 44% to 0%.

Data 2
For the 20 physicians that had all the required data elements and met the minimum number of quality reporting events (10) for inclusion in the reliability analysis, there were a total of 2,230 quality reporting events. 319 of the events were considered exceptions for an exception rate of 0.14. The average number of exceptions for 20 physicians included is 15.95. The range of exception rates for physicians included 85% to 1%.

Data 1
For the 237 physicians that had all the required data elements and met the minimum number of quality reporting events (10) for inclusion in the reliability analysis, there were a total of 5,445 quality reporting events. 695 of the events were considered exceptions for an exception rate of 10.0%. The average number of exceptions for 237 physicians included is 3. The range of exception rates for physicians included 73.0% to 0%.

Data 2
For the 2,666 physicians that had all the required data elements and met the minimum number of quality reporting events (10) for inclusion in the reliability analysis, there were a total of 229,209 quality reporting events. 29,407 of the events were considered exception for an exception rate of 11.0%. The average number of exceptions for 2,666 physicians included is 11. The range of exception rates for physicians included 64.0% to 0%.

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*)

The rates of exceptions are consistent with research that has suggested that approximately 25% of patients undergoing colonoscopy have poor bowel preparation.(1, 2)

Reference

1. Van Dongen M. Enhancing bowel preparation for colonoscopy: an integrative review. Gastroenterol Nurs. 2012 Jan;35(1):36-44.

2. Lebwohl B, Wang TC, Neugut AI. Socioeconomic and other predictors of colonoscopy preparation quality. Dig Dis Sci. 2010 Jul;55(7):2014-20. Epub 2010 Jan 16.

As noted in recent recommendations from the US Multi-Society Task Force on Colorectal Cancer on Optimizing Adequacy of Bowel Cleansing for Colonoscopy, up to 20–25% of all colonoscopies are reported to have an inadequate bowel preparation. The rates of exceptions found in the data analysis are fairly consistent with this research and are necessary to account for those situations when it is medically appropriate for a patient to have had a recommended follow-up interval of less 10 years for repeat colonoscopy. Without exceptions, the performance rate would not accurately reflect the true performance of the reporting physician. This would result in an increase in performance failures and false negatives.

Johnson DA, Barkun AN, Cohen LB, et al; US Multi-Society Task Force on Colorectal Cancer. Optimizing adequacy of bowel cleansing for colonoscopy: recommendations from the US multi-society task force on colorectal cancer. Gastroenterology. 2014 Oct;147(4):903-24. doi: 10.1053/j.gastro.2014.07.002._____

2b3. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES

If not an intermediate or health outcome, or PRO-PM, or resource use measure, skip to section 2b4.

2b3.1. What method of controlling for differences in case mix is used?

☒ **No risk adjustment or stratification**

☐ **Statistical risk model with** Click here to enter number of factors **risk factors**

☐ **Stratification by** Click here to enter number of categories **risk categories**

☐ **Other,** Click here to enter description

2b3.1.1 If using a statistical risk model, provide detailed risk model specifications, including the risk model method, risk factors, coefficients, equations, codes with descriptors, and definitions.

Not applicable

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.

Not applicable

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?

Not applicable

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?

Not applicable

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.**

Not applicable

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*)

Not applicable

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*)**:**

Not applicable

2b3.7. Statistical Risk Model Calibration Statistics (*e.g., Hosmer-Lemeshow statistic*):

Not applicable

2b3.8. Statistical Risk Model Calibration – Risk decile plots or calibration curves:

Not applicable

2b3.9. Results of Risk Stratification Analysis:

Not applicable

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*)

Not applicable

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*)

Not applicable

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)*

Measures of central tendency, variability, and dispersion were calculated.

Measures of central tendency, variability, and dispersion were calculated.

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*)

Data 1
Based on the sample of 177 included physicians, the mean performance rate is 0.5343, the median performance rate is 0.64 and the mode is 0.0. The standard deviation is 0.31 The range of the performance rate is 1.0, with a minimum rate of 0.00 and a maximum rate of 1.00. The interquartile range is 0.48. The 75th percentile is 0.78 and the 25th percentile is 0.3.

Data 2

Based on the sample of 20 included physicians, the mean performance rate is 0.3148, the median performance rate is 0.24 and the mode is 0. The standard deviation is 0.34 The range of the performance rate is 0.89, with a minimum rate of 0.00 and a maximum rate of 0.89. The interquartile range is 0.68. The 75th percentile is 0.68 and the 25th percentile is 0.0.

Data 1

Based on the sample of 237 included physicians, the mean performance rate is 0.93 the median performance rate is 1.00 and the mode is 1.0. The standard deviation is 0.15. The range of the performance rate is 0.95, with a minimum rate of 0.05 and a maximum rate of 1.0. The interquartile range is 0.09 (1.0–0.91).

Data 2

Based on the sample of 2,666 included physicians, the mean performance rate is 0.88 the median performance rate is 0.94 and the mode is 1.0. The standard deviation is 0.17. The range of the performance rate is 0.99, with a minimum rate of 0.01 and a maximum rate of 1.0. The interquartile range is 0.12 (.98–0.86).

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?*)

Data 1

The range of performance from 0.00 to 1.00 suggests there's clinically meaningful variation across physicians' performance.

Data 2

The range of performance from 0.00 to 0.89 suggests there's clinically meaningful variation across physicians' performance.

Data 1

The range of performance from 0.05 to 1.0 suggests there's clinically meaningful variation across physicians' performance.

Data 2

The range of performance from 0.01 to 1.0 suggests there's clinically meaningful variation across physicians' performance.

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*)

This test was not performed for this measure.

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*)

This test was not performed for this measure.

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*)

This test was not performed for this measure.

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*)

The PQRS dataset provided to us by CMS did not contain missing data so this test was not performed. Nevertheless, missing data may have been rejected when submitted to CMS in which case those values would not be counted towards measure performance. There is no indication that this missing data was systematic, thus their omission would lead to unbiased performance results.

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*)

This test was not performed for this measure. There was no missing data.

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*)

The PQRS dataset provided to us by CMS did not contain missing data so this test was not performed. Nevertheless, missing data may have been rejected when submitted to CMS in which case those values would not be counted towards measure performance. There is no indication that this missing data was systematic, thus their omission would lead to unbiased performance results.

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.

Coded by someone other than person obtaining original information (e.g., DRG, ICD-9 codes on claims)

If other:

3b. Electronic Sources

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

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

ALL data elements are in defined fields in electronic health records (EHRs)

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

N/A

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

Attachment:

3c. Data Collection Strategy

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

3c.1. Required for maintenance of endorsement. Describe difficulties (as a result of testing and/or operational use of the measure) regarding data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, time and cost of data collection, other feasibility/implementation issues.

IF instrument-based, consider implications for both individuals providing data (patients, service recipients, respondents) and those whose performance is being measured.

This measure was found to be reliable and feasible for implementation.

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

N/A

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 Payment Program https://qpp.cms.gov/ Payment Program Quality Payment Program https://qpp.cms.gov/

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

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

Current Use 1

The GI Quality Improvement Consortium, Ltd. ("GIQuIC") is an educational and scientific 501(c)(3) organization established by gastroenterologists, physicians specializing in digestive disorders. GIQuIC is a joint initiative of the American College of Gastroenterology (ACG) and the American Society for Gastrointestinal Endoscopy (ASGE). GIQuIC is a procedure-focused benchmarking registry using established quality indicators. The geographic area is the entire United States. GIQuIC registry participants have contributed real-time procedure related data from over 100,000 colonoscopies, not claims data, and the growth rate for the registry has increased to almost 2,000 new cases per week in recent months, with an accompanying surge in the growth of the number of practices involved in this quality improvement effort. GIQuIC is a national registry that fosters the ability of endoscopists and endoscopy facilities to benchmark themselves, and provides impetus for quality improvement. Some 84 data fields for colonoscopy are collected and ten quality measures are benchmarked, including rate of cecal intubation, adenoma detection rate, prep assessment, and appropriate indications for procedure, among others. Currently, hundreds of physicians from endoscopy centers nationwide have registered to participate in this ground-breaking initiative.

http://giquic.gi.org/

Current Use 2

Multiple QCDRs: Able Health, Academic Research for Clinical Outcomes (ARCO) in collaboration with ReportingMD, Inc; Citiustech, Inc.; Health-Advanta; Meditab Software, Inc.; Med-Xpress Registry; New Hampshire Colonoscopy Registry, Searfoss Consulting Group, Sovereign QCDR Registry.

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

This measure is currently publicly reported in the Quality Payment Program as a high-priority measure and has been reported in PQRS since 2009.

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.*)

This measure is currently publicly reported in the Quality Payment Program as a high-priority measure and has been reported in PQRS since 2009.

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.

Measure has been implemented in the Quality Payment Program (QPP) as an individual measure for claims and registry reporting, feedback is provided via CMS QRUR reports. Measure is also implemented in multiple QCDRs where feedback is required quarterly.

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.

Measure has been implemented in the Quality Payment Program (QPP) as an individual measure for claims and registry reporting, feedback is provided via CMS QRUR reports. Measure is also implemented in multiple QCDRs where feedback is required quarterly.

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.

Feedback was received to include information about familial or personal history of colonic polyps as well as life expectancy of patient. Measure was modified to include in denominator exception in 2016.

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

No feedback obtained from those being measured

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

No feedback obtained from other users.

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.

Measure was reviewed with ASC contractor for the parallel measure in the ASC program, physican experts, and all three GI societies, and from CMS. Consensus was reached and measure was modified to include in denominator exception.

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.

Performance measurement serves as an important component in a quality improvement strategy but performance measurement alone will not achieve the desired goal of improving patient care. Measures can have their greatest effect when they are used judiciously and linked directly to operational steps that clinicians, patients, and health plans can apply in practice to improve care.

4b2. Unintended Consequences

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

4b2.1. Please explain any unexpected findings (positive or negative) during implementation of this measure including unintended impacts on patients.

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

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

We are not aware of any unexpected benefits from implementation of this measure.

5. Comparison to Related or Competing Measures

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

5. Relation to Other NQF-endorsed Measures

Are there related measures (conceptually, either same measure focus or target population) or competing measures (conceptually both the same measure focus and same target population)? If yes, list the NQF # and title of all related and/or competing measures.

Yes

5.1a. List of related or competing measures (selected from NQF-endorsed measures)

0572 : Follow-up after initial diagnosis and treatment of colorectal cancer: colonoscopy

0659 : Colonoscopy Interval for Patients with a History of Adenomatous Polyps- Avoidance of Inappropriate Use

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

ASC-9: Appropriate Follow-up Interval for Normal Colonoscopy in Average Risk Patients - Telligen

ASC-10: Colonoscopy Interval for Patients with a History of Adenomatous Polyps – Avoidance of Inappropriate Use - Telligen

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?
No
5a.2. If the measure specifications are not completely harmonized, identify the differences, rationale, and impact on interpretability and data collection burden.
The list of measures above, includes several different populations and capture different elements in the numerator. None of them are aiming to capture the same information as measure 0658. Measures 0572, ACP-018-10, and 0392 actually aim to capture specific elements within the colonoscopy report or pathology report (after colon/rectum resection). Measure 0034 intends to capture one of four different types of colorectal cancer screening tests, instead of looking specifically at the interval between colonoscopies. Measure 0659 focuses on a different patient population, as the patients in 0659 have had a history of a prior colonic polyp(s) in previous colonoscopy findings. The patient population in measure 0659 has a different follow up interval recommendation, according to evidence based guidelines.
5b. Competing Measures
The measure is superior to competing measures (e.g., is a more valid or efficient way to measure);
OR
Multiple measures are justified.
5b.1. If this measure conceptually addresses both the same measure focus and the same target population as NQF-endorsed measure(s):
Describe why this measure is superior to competing measures (e.g., a more valid or efficient way to measure quality); OR provide a rationale for the additive value of endorsing an additional measure. (Provide analyses when possible.)
There are no competing measures.

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.

Attachment Attachment: NQF_GI_GUProject_Stage2_Checklist_Memo_AMAPCPI_Answers-634935166469631059-636426432400364692.docx

Contact Information

Co.1 Measure Steward (Intellectual Property Owner): American Gastroenterological Association

Co.2 Point of Contact: David, Godzina, dgodzina@gastro.org, 301-272-1600-

Co.3 Measure Developer if different from Measure Steward: American Gastroenterological Association

Co.4 Point of Contact: David, Godzina, dgodzina@gastro.org, 301-272-1600-

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.

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

Co-chairs

John Allen, MD, MBA, AGAF (Gastroenterology)

Doug Faigel, MD (Gastroenterology)

Work Group Members

Nancy Baxter, MD, PhD, FACRS, FACS (Colon and Rectal Surgery)

Stephen Bickston, MD, AGAF (Gastroenterology)

Joel V. Brill, MD, AGAF, FASGE, FACG, CHCQM (Gastroenterology)

Kirk Brandon, MBA (Business Administration/Coding)

Jason A. Dominitz, MD, MHS, AGAF (Gastroenterology)

Ira L. Flax, MD, FACG (Gastroenterology)

Karen E. Hall, MD, PhD (Geriatrics)

Robert Haskey, MD, FACS (General Surgery, Health Plan representative)

Brian C. Jacobson, MD, MPH (Gastroenterology)

David Lieberman, MD (Gastroenterology)

Klaus Mergener, MD, PhD, CPE, FACP, FACG, FASGE, FACPE (Gastroenterology)

Bret Petersen, MD, FASGE (Gastroenterology)

Irving M. Pike, MD, FACG (Gastroenterology)

Bart Pope, MD (Family Medicine)

Harry Sarles, MD, FACG (Gastroenterology)

Kay Schwebke, MD, MPH (Specialty: Internal Medicine, Infectious Diseases & Medical Informatics)

Tom Lynn, MD (Medical Informatics, Methodology)

Emily E. Volk, MD, FCAP (Pathology)

Michael Weinstein, MD Specialty: Gastroenterology)

American Gastroenterological Association

Debbie Robin, MSN, RN, CHCQM

American Society for Gastrointestinal Endoscopy

Jill Blim

Chris Recker, RN, MPH

Martha Espronceda

American College of Gastroenterology

Julie Cantor-Weinberg, MPP

American Medical Association

Joseph Gave, MPH

Karen Kmetik, PhD

Shannon Sims, MD, PhD

Beth Tapper, MA

Consortium Consultants

Rebecca Kresowik

Timothy Kresowik, MD

Measure Developer/Steward Updates and Ongoing Maintenance

Ad.2 Year the measure was first released: 2008

Ad.3 Month and Year of most recent revision: 08, 2008

Ad.4 What is your frequency for review/update of this measure? See Ad.9.

Ad.5 When is the next scheduled review/update for this measure? 08, 2013

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

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

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

The AMA's, PCPI's and National Committee for Quality Assurance's significant past efforts and contributions to the development and updating of the Measures is acknowledged. AGA, ASGE and ACG are solely responsible for the review and enhancement ("Maintenance") of the Measures as of August 14, 2014.

AGA, ASGE and ACG encourage use of the Measures by other health care professionals, where appropriate.

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

 For the Merit-Based Incentive Payment System, American Gastroenterological Association is the primary steward for measure revisions.

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

Ad.7 Disclaimers: N/A

Ad.8 Additional Information/Comments: Coding/Specifications updates occur annually. The PCPI has a formal measurement review process that stipulates regular (usually on a three-year cycle, when feasible) review of

the measures. The process can also be activated if there is a major change in scientific evidence, results from testing or other issues are noted that materially affect the integrity of the measure.