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 #: 1859

Corresponding Measures:

De.2. Measure Title: RAS gene mutation testing performed for patients with metastatic colorectal cancer who receive anti-epidermal growth factor receptor monoclonal antibody therapy

Co.1.1. Measure Steward: American Society of Clinical Oncology

De.3. Brief Description of Measure: Percentage of adult patients (aged 18 and over) with metastatic colorectal cancer who receive anti-epidermal growth factor receptor monoclonal antibody therapy for whom RAS (KRAS and NRAS) gene mutation testing was performed

1b.1. Developer Rationale: We envision that use of this measure will improve concordance with recommendations for expanded RAS testing. Evidence now supports testing for NRAS in addition to KRAS mutations. ASCO anticipates a greater performance gap due to the guideline update, which is a relatively new requirement in the field. Clinical trials data show that the benefit of using EGFR inhibitors in treating metastatic colorectal cancer, either as monotherapy or in combination with other treatment regimens, is limited to non-existent in patients with RAS-mutated tumors. These data strongly suggest that patients with RAS mutations are better served with other therapies, especially considering the harms and costs of anti-EGFR treatment.

S.4. Numerator Statement: RAS (KRAS and NRAS) gene mutation testing performed prior to initiation of anti-EGFR monoclonal antibody therapy

S.6. Denominator Statement: Adult patients with metastatic colorectal cancer who receive anti-EGFR monoclonal antibody therapy

S.8. Denominator Exclusions: None

De.1. Measure Type: Process

S.17. Data Source: Paper Medical Records, Registry Data

S.20. Level of Analysis: Clinician : Group/Practice

IF Endorsement Maintenance – Original Endorsement Date: Oct 22, 2012 **Most Recent Endorsement Date:** Oct 22, 2012

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

Preliminary Analysis: Maintenance of Endorsement

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

Criteria 1: Importance to Measure and Report

1a. Evidence

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

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

The developer provides the following evidence for this measure:

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

Summary of prior review in 2016

- The developer provided evidence that was based on the following clinical practice guideline:
    - Based on a systematic review of the relevant literature, all patients with metastatic colorectal carcinoma who are candidates for anti-EGFR monoclonal antibody therapy should have their tumor tested for KRAS mutations in a CLIA-accredited laboratory.
        - No grading criteria was provided, however the guideline was approved by a unanimous vote by a panel that was selected and charged by the ASCO Health Services Committee.

Changes to evidence from last review

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

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

Updates:

- The updated evidence for this measure was based on the following clinical practice guidelines or recommendations:
    - ASCO recommendation: Colorectal carcinoma patients being considered for anti-EGFR therapy must receive RAS mutational testing. Mutational analysis should include KRAS and NRAS codons 12, 13 of exon 2; 59, 61 of exon 3; and 117 and 146 of exon 4 ("expanded" or "extended" RAS). **Evidence strength**: Convincing/Adequate; **Evidence quality**: High/Intermediate; **Recommendation grade**: Expert consensus opinion.
        - The developer notes serious limitations in the recommendation grade, such as limited strength of evidence, intermediate-to-low quality of evidence, and balance of benefits and harms, values, or costs.

- NCCN guideline on colon cancer: All patients with metastatic colorectal cancer should have tumor tissue genotyped for RAS (KRAS and NRAS) and BRAF mutations individually or as part of a next-generation sequencing (NGS) panel. Patients with any known KRAS mutation (exon 2, 3, 4) or NRAS mutation (exon 2, 3, 4) should not be treated with either cetuximab or panitumumab. […] A sizeable body of literature has shown that tumors with a mutation in codon 12 or 13 of exon 2 of the KRAS gene are essentially insensitive to cetuximab or panitumumab therapy… More recent evidence shows mutations in KRAS outside of exon 2 and mutations in NRAS are also predictive for a lack of benefit of anti-EGFR therapies. **Evidence strength/quality and recommendation grade**: the intervention is appropriate based on lower-level evidence.
    - The guidelines do not present evidence used for the recommendation specific to RAS mutation status; however, evidence is provided on the benefits and harms of EGFR inhibitors.

Questions for the Committee:

- For structure, process, and intermediate outcome measures:
    - What is the relationship of this measure to patient outcomes?
    - How strong is the evidence for this relationship?
    - Is the evidence directly applicable to the process of care being measured?

Guidance from the Evidence Algorithm

Evidence is based on expert opinion and is about EGFR inhibitors, as opposed to RAS mutation status (Box 3) → empirical evidence is submitted (Box 7) → Empirical evidence includes all studies (Box 8) → Evidence indicates moderate certainty that benefits outweigh undesirable effects (Box 9)

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

RATIONALE: The evidence presents serious limitations in strength and quality, and does not address what is being measured specifically (RAS mutation status).

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

Maintenance measures – increased emphasis on gap and variation

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

- The developer used the following 2017 MIPS performance registry data provided from CMS:

  Number of unique entities:

  Frequency

  43

  Denominators

Min	Q1	Median	Mean	Q3	Max
1	2	6	11.51	12	82

  Measure Distribution:

Min	Q1	Median	Mean	Q3	Max	CI for mean	Percent outside CI
0	0.9902	1	0.9123	1	1	(0.85, 0.98)	95.35

- 54% of practices perform at 100%, however many practices perform at rates ranging from 0% to 76%, indicating room for improvement.

- The following are data presented for practices:
  - Mean = 76.1%, confidence interval (0.65, 0.87)
  - Practice min = 0%
  - Practice max = 100%
  - Practice perfect outside confidence interval = 80.49%
- The developer also presented a summary of data from the literature that indicates performance gaps or overall less than optimal performance.

Disparities
No disparities data was presented. However, the developer cited a 2017 SEER study that found overall proportion of KRAS testing was only 22.7% among the sample population with variation by geographic region and patient characteristics, indicating disparities in KRAS testing.

Questions for the Committee:

- Is there a gap in care that warrants a national performance measure?
- If no disparities information is provided, are you aware of evidence that disparities exist in this area of healthcare?

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

RATIONALE:

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

1a. Evidence to Support Measure Focus: For all measures (structure, process, outcome, patient-reported structure/process), empirical data are required. How does the evidence relate to the specific structure, process, or outcome being measured? Does it apply directly or is it tangential? How does the structure, process, or outcome relate to desired outcomes? For maintenance measures –are you aware of any new studies/information that changes the evidence base for this measure that has not been cited in the submission? For measures derived from a patient report: Measures derived from a patient report must demonstrate that the target population values the measured outcome, process, or structure."

- This is a process measure that is intended to assess the clinician for having RAS testing result available prior to ordering EGFR antibody. RAS results are predictive of effectiveness of EGFR antibodies and an important clinical data as supported by many studies.
- Evidence based
- Evidence has been stronger for link between RAS testing and EGFR treatment; failure to test may affect choice of treatment inappropriately
- High level of evidence supporting lack of utility of EGFR inhibitors wehn RAS is mutated.
- Seems reasonable assumption
- There is good evidence for support of this measure. The challenge is denominator: "Adult patients with metastatic colorectal cancer who have a RAS (KRAS or NRAS) gene mutation". There is no standard for testing for RAS mutation and thus a portion of patients will have unknown RAS and excluded from this measurement. The greater threat is that those patient could have received EGFR antibodies should testing be available and yet will be excluded from a beneficial agent. The point is

that mandating testing for RAS at the benining of diagnosis of mCRC is more important than the currently reported measure.

- The measure is intended to improve concordance with recommendations for expanded RAS testing in patients prior to the initiation of EGFR inhibitors in metastatic colorectal cancer. This is a process measure. The medical evidence strongly suggests that patients with RAS mutations do not benefit from anti-EGFR treatment. The assumption is that if this testing is done, then targeted agents will be used appropriately. The available evidence is not directly related to this measure but this clinical information is a required process in order to direct therapy. The guideline that patients undergo RAS testing and the specific mutations to be tested prior to the initiation of therapy is recommended by multiple entities including ASCO and NCCN. The evidence for the avoidance of EGFR in patients with RAS mutations is of high quality and quantity and the results are consistent across multiple studies. The evidence cited by the NCCN was graded as Level 2A, lower quality evidence but uniform consensus to support the guideline. Although the direct evidence for this measure is rated as low, I believe that the evidence to support the appropriate therapeutic interventions using test results is clinically relevant and important to measure.
- The measure seeks to avoid the use of monoclonal therapies in patients with metastatic colon cancer with RAS mutations as multiple studies have shown that these patients do not benefit from these types of therapies. It is a process measure and an appropriate use measure with a goal to avoid the use of targeted therapies in patients without the target. The evidence for the practice recommendation is associated with high level evidence and is directly linked to the process. The guideline is supported by multiple entities. ASCO's recommendation is based on 34 studies, including 29 systematic studies, two meta-analyses, one randomized controlled trial, one prospective cohort study and one retrospective cohort study. The evidence is of high quality and quantity and the results are consistent across multiple studies. The developers updated the evidence for this submission and focused on the subtypes of KRAS mutations to further identify patients unlikely to benefit from specific therapy. The evidence cited by the NCCN was graded as Level 2A, lower quality evidence but uniform consensus to support the guideline. This is also a guideline from the American College for Clinical Pathology. The structure of the measure is related to the desired outcome.
- Evidence does not match the metric which is being evaluated fully.
- Overall level of evidence is low. Systemtatic reviews are presented but much of the evidence is lower level.
- It appears the evidence does not address the what is being measured (as noted in the preliminary analysis).

1b. Performance Gap: Was current performance data on the measure provided? How does it demonstrate a gap in care (variability or overall less than optimal performance) to warrant a national performance measure? Disparities: Was data on the measure by population subgroups provided? How does it demonstrate disparities in the care?

- a performance gap from the analysis of 2017 MIPS data was provided. The data is presented per practice with a mean of 76%. However, the range of the data (0-100) suggests that there is huge disparity in clinicians practice and this measure needs to be done per provider. This in turn create the number issue that may make the results uninterpretable.
- There is a gap
- yes, there is a gap that is signficiant

- Although half of practices are performing at 100%, the remainder range from0-76%, indicating that a performance gap remains with possible geographic disparities.
- Large Gap
- The mean performance of the group is 91%, statistically different from 100% and thus justifying they measure. what is unknown is how many patients had RAS testing that was not captured as part of this measure.
- The developer provided 2017 MIPS performance from registry data provided from CMS. 54% of practices perform at 100%, however many practices perform at rates ranging from 0% to 76%, indicating room for improvement. The measure developer did not provide disparities data and noted that while this measure is included in the MIPS program, this program has not yet made disparities data available for ASCO to analyze the report. The developers also presented data from the literature to support a performance gap.
- The developer provided 2017 MIPS performance from registry data provided from CMS. The 2017 data was from 158 providers representing 43 practices and 495 individual patients. While the majority (approximately 76.7%) of practices performed at 100% with a mean performance of 91%, the mean performance rate of 91% is statistically significant from 100% suggesting room for improvement remains across practices. The measure developer did not provide disparities data and noted that while this measure is included in the MIPS program, this program has not yet made disparities data available for ASCO to analyze the report.
- There is a performance gap and measure has room for improvement. Data was run to look at multiple disparities data and showed some potentially meaningful differences.
- A performance gap is clearly present with opportunities for improvement in care with overall mean of 76% for practices with range of 0% to 100%
- I agree with the preliminary rating of "High".

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability: Specifications and Testing

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

2c. For composite measures: empirical analysis support composite approach

Reliability

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

2a2. Reliability testing demonstrates if the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise enough to distinguish differences in performance across providers. For maintenance measures – less emphasis if no new testing data provided.

Validity

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

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

Composite measures only:

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

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

Evaluators: NQF Staff

Scientific Acceptability: Preliminary Analysis

Measure Number: 1859

Measure Title: RAS gene mutation testing performed for patients with metastatic colorectal cancer who receive anti-epidermal growth factor receptor monoclonal antibody therapy

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 developer conducted this testing at the facility level but indicated that level of analysis is group/practice. The developer should resubmit testing at the appropriate level of analysis.

RELIABILITY: TESTING

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

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

   ☐ **Yes** ☐ **No**
6. **Assess the method(s) used for reliability testing**

   Submission document: Testing attachment, section 2a2.2

   - The developer computed signal-to-nnoise scores to address precision of measurement (measure score) and used a beta-binomial model.
   - The developer conducted this testing at the facility level but indicated that level of analysis is group/practice. The developer should resubmit testing at the appropriate level of analysis.
   - The developer indicated critical data element testing but did not report data element reliability (2a2.1 on the testing form).
7. **Assess the results of reliability testing**

   Submission document: Testing attachment, section 2a2.3

   - A reliability of zero implies that the variability in the measure is attributed to measurement error, while a reliability of one implies that the variability is attributable to real differences in facility performance. 0.70 – 0.80 reliability is considered an acceptable threshold. 0.80 – 0.90 is considered high reliability. And 0.90 – 1.00 is considered very high.
   - The developers reported a mean reliability of 0.8908 which is considered very high according to Adams' definition.

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.**

Per Box 2 of the reliability algorithm, testing does not match measure specifications, i.e. level of analysis. The developer reports facility-level testing but indicates that this measure be specified at the group/practice level of analysis.

VALIDITY: ASSESSMENT OF THREATS TO VALIDITY

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

Submission document: Testing attachment, section 2b2.

- No exclusions for this measure.

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

Submission document: Testing attachment, section 2b4.

- Small sample size (denominator median = 3) may be impacting the presented results.

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.

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

Submission document: Testing attachment, section 2b6.

16. **Risk Adjustment**

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

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

☐ Yes ☐ No ☒ Not applicable

16c. **Social risk adjustment:**

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

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

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

16d.**Risk adjustment summary:**

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

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

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

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

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

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

For cost/resource use measures ONLY:

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

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

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

VALIDITY: TESTING

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

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

    ☐ **Face validity**

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

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

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

    Submission document: Testing attachment, section 2b2.2

    - The developer performed a Pearson correlation analysis to determine the association between the performance scores of the shared providers.
    - The developer interpreted the correlation scores in the following way:
        - > 0.40 correlation coefficient = strong correlation
        - 0.20 – 0.40 correlation coefficient = moderate correlation
        - < 0.20 correlation coefficient = weak coefficient

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

    Submission document: Testing attachment, section 2b2.3

    - The correlation was 0.49, indicating a strong, positive correlation between performance scores of the shared providers.

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

    Submission document: Testing attachment, section 2b1.

    ☒ **Yes**

    ☐ **No**

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

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

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

    ☐ **Yes**

    ☐ **No**

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

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

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

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

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

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

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

    Developer assessed all applicable threats to validity (statistically significant and meaningful differences, missing data/nonresponse) (Box 1) → Pearson correlation analysis results were provided (Box 2) → Validity testing conducted at the TIN level (Box 6) → 0.49 correlation coefficient, which is considered strong (Box 7a) → HIGH, highest possible rating is high.

ADDITIONAL RECOMMENDATIONS

27. **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: Which data elements, if any, are not clearly defined? Which codes with descriptors, if any, are not provided? Which steps, if any, in the logic or calculation algorithm or other specifications (e.g., risk/case-mix adjustment, survey/sampling instructions) are not clear? What concerns do you have about the likelihood that this measure can be consistently implemented?

- not sure.
- No significant concerns
- uncertain since not all testing is done at same location, but likely to be achievable
- testing level of analysis facility rather than group. no concerns about reliability
- none
- The data elements are derived from medical records (paper?), this reliability of the data is major question. Unknown and undocumented RAS status is a reliability threat to this measure.
- The data elements are clearly defined. The measure description is complete and concise. The developer submitted updates to the measure specifications to include NRAS and KRAS testing which were included in 2018. I believe that this measure can be consistently implemented.
- The data elements are clearly defined. The measure description is complete and concise. The developer submitted updates to the measure specifications to include NRAS and KRAS testing which were included in 2018. I believe that this measure can be consistently implemented.
- Most of the data elements are clearly defined. The only concern would be the pitfalls of manually abstracted measurement which could introduce different data intepretations.
- Developer reports as very high with mean observed reliablity of 95%, but am not finding satisfactory clarification of the level of analysis contradiction (group/practice versus facility).
- I agree the appropriate level of analysis should be submitted.

2a2. Reliability - Testing: Do you have any concerns about the reliability of the measure?

Yes, the testing is based on old data and may not reflect new data and more importantly newer methods of documentation and testing

- No major concern
- minimal
- No
- no
- Yes, see response to the above question.
- The NQF staff noted that the developer submitted testing at the facility level but the testing is reported to be at the group/practice level. Testing was conducted to measure the ratio of signal to noise testing. The data was abstracted from the medical record or tumor registry data. The datasets used for testing were from 2017 MIPS data. For the 2019 submission, the 2017 data was from 158 providers representing 43 practices and 495 individual patients. Signal to noise analysis yielded a reliability greater than 0.90 which is considered very high. The mean reliability of 95% observed is categorized as high reliability and the 10th percentile is 74%. The NQF staff rates the reliability testing as insufficient but I would clarify the testing level with the developers since QOPI does its analyses at the group/practice level. The level of testing should be clarified before a final consideration. Otherwise, I would rate the reliability as high.
- I have no concerns. Testing was conducted to measure the ratio of signal to noise testing. The testing met the NQF criteria for high since the testing was done at the measure score level. The data was abstracted from the medical record or tumor registry data. The datasets used for testing were 2011 QOPI data and 2017 MIPS data, which are consistent with the measure specifications. For the 2019 submission, the 2017 MIPs data was from 158 providers representing 43 practices and 495 individual patients. Facility level reliability testing was found to be a mean of 0.9465 which is associated with a high level of reliability.
- No
- Yes, need clarification of level of analysis for reliability testing.
- I agree with the preliminary analysis of "Insufficient".

2b1. Validity -Testing: Do you have any concerns with the testing results?

- No
- It was acceptable
- minimal
- no
- None
- I don't see any validity issue.
- I have no concerns. Testing was done on the performance measure score. The developer performed a Pearson correlation analysis to determine the association between the performance scores of the shared providers. The correlation coefficient observed was 0.49 indicating a strong correlation.
- I have no concerns. A correlation analysis was completed using 2017 MIPS data. KRAS gene mutation testing performed for patients with metastatic colorectal cancer who receive anti-epidermal growth factor receptor monoclonal antibody therapy (QI #451/ NQF#1859) was correlated to this measure since the population is similar. This seems to be a reasonable measure to

correlate the results. Testing was done on the performance measure score. The correlation coefficient observed was 0.49 based on 28 matching practices indicating a strong correlation.

- Yes
- No concerns high validity demonstrated
- I agree with the preliminary rating of "High".

2b4-7. Threats to Validity (Statistically Significant Differences, Multiple Data Sources, Missing Data)2b4. Meaningful Differences: How do analyses indicate this measure identifies meaningful differences about quality? 2b5. Comparability of performance scores: If multiple sets of specifications: Do analyses indicate they produce comparable results? 2b6. Missing data/no response: Does missing data constitute a threat to the validity of this measure?

- Missing data is always a threat to validity for paper based measures
- No response or missing data could be a problem
- missing data may be an issue, but unlikely to have a systemic problem
- None appreciated.
- No concerns
- The likelihood of any validity threat is very low.
- There are no threats to the validity. There were no risk adjustments and there were no exclusions to testing. The testing appears to support the ability to detect meaningful differences. The developers defined a meaningful difference as the presence of a significant spread between the minimum and maximum scores or a significant spread between median and either the minimum or maximum scores. They presented data from 2017 MIPS reporting at the practice and individual that suggested the ability to detect meaningful differences and indicated the opportunity for improvement in performance. Performance data from MIPS data 2017 does not include data for expanded RAS testing as those changes were implemented in 2018. There was no missing data in the MIPs data.
- There are no threats to the validity. There were no risk adjustments and there were no exclusions to testing. The testing appears to support the ability to detect meaningful differences. The developers defined a meaningful difference as the presence of a significant spread between the minimum and maximum scores or a significant spread between median and either the minimum or maximum scores. Testing from the 2017 MIPS reporting demonstrated a practice mean of 91.23% with a confidence interval (0.85, 0.98). The range was 0% to 100% and the practice percent outside confidence interval was 93.35%. At the individual level, the mean was 91.7% with a confidence interval (0.88, 0.95) and a range of 0% to 100%. The individual clinician percent outside confidence interval was 100%. A majority (approximately 76.7%) of practices performed at 100%. Performance data from MIPS data 2017 does not include data for expanded RAS testing as those changes were implemented in 2018. There was no missing data in the MIPs data.
- N/A
- No concerns.
- No concerns

2b2-3. Other Threats to Validity (Exclusions, Risk Adjustment)2b2. Exclusions: Are the exclusions consistent with the evidence? Are any patients or patient groups inappropriately excluded from the measure?2b3. Risk Adjustment: If outcome (intermediate, health, or PRO-based) or resource use performance measure: Is there a conceptual relationship between potential social risk factor variables

and the measure focus? How well do social risk factor variables that were available and analyzed align with the conceptual description provided? Are all of the risk-adjustment variables present at the start of care (if not, do you agree with the rationale provided)? Was the risk adjustment (case-mix adjustment) appropriately developed and tested? Do analyses indicate acceptable results? Is an appropriate risk-adjustment strategy included in the measure?

- No other threats.
- No risk adjustment
- no concerns
- none appreciated. exclusions appropriate
- Yes all reasonable
- exclusions are consistent with evidence. However, the population with unknown RAS can not benefit from this measure.
- There are no exclusions. There is no risk adjustment.
- There were no exclusions and no risk adjustments.
- I am concerned about potential exclusions due to patient preference and patient cost
- No Concerns
- No concerns

Criterion 3. Feasibility

Maintenance measures – no change in emphasis – implementation issues may be more prominent

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

- Data is collected by and used by healthcare personnel during provision of care.
- Data is abstracted from records by someone other than the person collecting the data.
- Only some data elements are in defined fields in electronic sources.
- A licensing agreement is required prior to commercial use of this measure.
- This may be burdensome as it may require chart abstractions. Use of this measure through EHRs would lessen this burden. The developer reports that they are in the process of assessing feasibility of developing an eCQM.

Questions for the Committee:

- Are the required data elements routinely generated and used during care delivery?
- Are the required data elements available in electronic form, e.g., EHR or other electronic sources?

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

RATIONALE:

Committee Pre-evaluation Comments:
Criteria 3: Feasibility

3. Feasibility: Which of the required data elements are not routinely generated and used during care delivery? Which of the required data elements are not available in electronic form (e.g., EHR or other

electronic sources)? What are your concerns about how the data collection strategy can be put into operational use?

- some data elements are not available in EHR and are dependent upon data abstraction
- Not the most convenient but possible
- should be already recorded
- some fields not in EHR. moderate
- reasonably Feasible
- RAS testing may be done by an outside entity (other than the ones available on lab reports/medical records) and thus may not be available for evaluation.
- The data elements are those that are routinely generated during the course of care. Most of the data is not available in an electronic format. The data elements are collected by chart audit or through a cancer registry. The measure is already being used for QOPI reporting and MIPs and has been operational for years. I have no concerns about the feasibility of this measure.
- The data elements are routinely generated during the course of care. Many of the data elements are not available in an electronic format. This measure requires chart audit to complete the data sets. This measure has been operational for many years and has proven to be feasible.
- Elements are documented during routine care however they are either documented in a narrative note, an order (i.e. pain medication, referral), or in an electronic way depending on EHR build. There is no standard element built into most EHR platforms. This metric requires manual audit.
- Reported that some data elements are defined in EHRs but not finding clear documentation of the specific data elements that need to be manually abstracted.
- I agree with the preliminary rating of "Moderate".

Criterion 4: Usability and Use

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

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

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

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

Current uses of the measure

Publicly reported? ☐ **Yes** ☒ **No**

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

OR

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

Accountability program details

- The measure is used in several accountability programs, including:
    - Merit-based Incentive Payment System (MIPS)
    - Quality Oncology Practice Initiative (QOPI)
    - Core Quality Measure Collaborative's (CQMC) Medical Oncology Core Measure Set

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

- Those being measured and other users have been given an opportunity to provide feedback on the measure performance or implementation. No specific feedback has been received by the developer aside from the multi-disciplinary technical expert panel during the measure development and maintenance process. Because no specific feedback was received, the TEP did not consider external feedback during revision of measure specifications or implementation.

Additional Feedback:

***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**

RATIONALE:

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 reports a high performance rate, with approximately 54% of practices performing at 100%. However, multiple practices are still operating at 0%. Mean performance is at 76%, indicating room for improvement.
- MIPS 2017 perormance data does not include RAS testing guideline changes made in 2018. The developer anticipates a greater performance gap to be made due to this guideline update.

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 states that they are currently unaware of any untintended consequences and benefits related to the measure.

Potential harms

- None reported

Additional Feedback:

***Questions for the Committee**:*

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

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

RATIONALE:

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

4a1. Use - Accountability and Transparency: How is the measure being publicly reported? Are the performance results disclosed and available outside of the organizations or practices whose performance is measured? For maintenance measures - which accountability applications is the measure being used for? For new measures - if not in use at the time of initial endorsement, is a credible plan for implementation provided?4a2. Use - Feedback on the measure: Have those being measured been given performance results or data, as well as assistance with interpreting the measure results and data? Have those being measured or other users been given an opportunity to provide feedback on the measure performance or implementation? Has this feedback has been considered when changes are incorporated into the measure?

- measure is used in QOPI, CQMC, and MIPS
- Useful
- being used by accountability programs
- not publically reported but in use in accountability program. no specific feedback.
- Yes
- The measure is not publicly reports. It is used as part of MIPS and QOPI
- The measure is already in use for QOPI and MIPs. The results of the measure can be used to improve performance. The measure has been updated to reflect new evidence.
- This measure is in active use in both QOPI and MIPs. The results of the measure are readily interpretable and can be used to improve performance. The measure has been updated to include new, relevant information regarding testing and the applicability of the associated therapeutic choices.
- Measure is used in multiple reporting programs
- No concerns, publicly reported in MIPS, QQPI and CQMC
- I agree with the preliminary rating of "Pass".

4b1. Usability – Improvement: How can the performance results be used to further the goal of high-quality, efficient healthcare? If not in use for performance improvement at the time of initial endorsement, is a credible rationale provided that describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations?4b2. Usability – Benefits vs. harms: Describe any actual unintended consequences and note how you think the benefits of the measure outweigh them.

- Can encourage providers to order RAS testing. However, if a provider is seeing few cases per year there is no mechanism to enforce the practice
- No harm, it's beneficial
- data should drive more testing
- no percieved unintended consequences
- Measure should reduce harms of therapy that is not useful for the disease.
- This measure is usable for sparing patients with RAS mutated from getting EGFR antibodies. However, it is not impacting those with unknown RAS, this is an unmet need population and completely excluded from this measure.
- The performance results are important to improve therapeutic choices for patients with metastatic colon cancer. This is potentially associated with improved outcomes and decreased toxicity. This also assures the avoidance of inappropriate agents. The benefits of this measure outweigh any risks for the implementation of the measure. I cannot identify any unintended consequences.
- The benefit of using the appropriate therapy which could result in improved clinical outcome and decreased toxicity as well as avoidance of inappropriate therapy far outweighs any risks. I am not aware of any unintended consequences of using this measure.
- Use of the metric would depend on the center- over 50% of centers are high performing and would not find value in continuing to measure.
- Multiple practices are performing at lower levels indicating a potential to improve the quality of health care.
- I agree with the preliminary rating of "High".

Criterion 5: Related and Competing Measures

Related or competing measures
1860 Patients with metastatic colorectal cancer and RAS gene mutation spared treatment with anti-epidermal growth factor receptor monoclonal antibodies

Harmonization
The measure specifications are harmonized. The developer states that 1859 is a complementary measure to 1860, which addresses the inverse of the quality action captured in 1859.

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

5. Related and Competing: Are there any related and competing measures? If so, are any specifications that are not harmonized? Are there any additional steps needed for the measures to be harmonized?

- 1860
- No action needed
- no issues
- very closely related to 1860, which measure rate at which patients with RAS mutations "didn't get EGFR therapy." These two measures would best be combined as they mirror each other.
- None

- No competing measure.
- There is one related measure which appears to be harmonized.
- There is an associated measure which has likely been harmonized with this measure.
- measures are harmonized
- No competing measures.
- I agree this measure is harmonized with measure 1860.

Public and Member Comments

Comments and Member Support/Non-Support Submitted as of: 02/14/2020

- No comments received

ADDITIONAL RECOMMENDATIONS

28. **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.**

Developer Submission

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

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

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

1859_Evidence_MSF5.0_Data.doc,1859_nqf_evidence_attachment_11.23.19.docx

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

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

Yes

1a. Evidence (subcriterion 1a)

Measure Number (*if previously endorsed*)**:** 1859

Measure Title: RAS gene mutation testing performed for patients with metastatic colorectal cancer who receive anti-epidermal growth factor receptor monoclonal antibody therapy

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

Date of Submission: 11/12/19

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

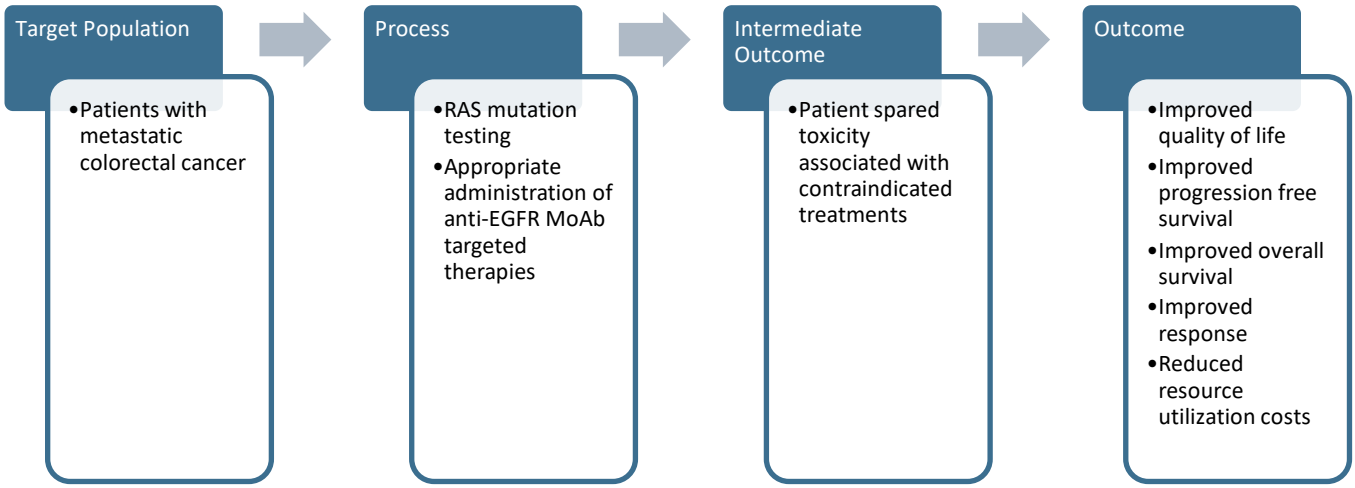
☒ Process: Click here to name what is being measured

☒ Appropriate use measure: Administration of anti-EGFR monoclonal antibody targeted therapies based on RAS mutation status

☐ 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.



The process evaluated in this measure is the completion of RAS testing to identify those patients who will not benefit from anti-EGFR therapy. Multiple studies, including a randomized controlled trial (RCT) support knowing a patient's tumor mutation status before consideration of use of an EGFR inhibitor in the treatment regimen.

The body of evidence addresses the relationship between RAS status in patients with metastatic colorectal cancer who underwent of anti-EGFR MoAB therapy, specifically cetuximab or panitumumab, and the outcomes of tumor response, progression-free survival, and overall survival. Patients with and without KRAS or NRAS mutations to exons 2, 3 or 4 who underwent anti-EGFR MoAb therapy were evaluated with respect to these outcomes in both single-arm and randomized trials. Additionally, this measure is directly supported by recommendations in American Society for Clinical Pathology, College of American Pathologists, Association for Molecular Pathology, and the American Society of Clinical Oncology and NCCN clinical practice guidelines.

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

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

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

1a.3. SYSTEMATIC REVIEW(SR) OF THE EVIDENCE (for INTERMEDIATE OUTCOME, PROCESS, OR STRUCTURE PERFORMANCE MEASURES, INCLUDING THOSE THAT ARE INSTRUMENT-BASED) If the

evidence is not based on a systematic review go to section 1a.4) If you wish to include more than one systematic review, add additional tables.

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

☒Clinical Practice Guideline recommendation (with evidence review)

- ☐ US Preventive Services Task Force Recommendation
- ☐ Other systematic review and grading of the body of evidence (*e.g., Cochrane Collaboration, AHRQ Evidence Practice Center*)
- ☐ Other

Source of Systematic Review: • **Title** • **Author** • **Date** • **Citation, including page number** • **URL**	Molecular Biomarkers for the Evaluation of Colorectal Cancer Antonia R. Sepulveda, Stanley R. Hamilton, Carmen J. Allegra, Wayne Grody, Allison M. Cushman-Vokoun, William K. Funkhouser, Scott E. Kopetz, Christopher Lieu, Noralane M. Lindor, Bruce D. Minsky, Federico A. Monzon, Daniel J. Sargent† Veena M. Singh, Joseph Willis, Jennifer Clark, Carol Colasacco, R. Bryan Rumble, Robyn Temple-Smolkin, Christina B. Ventura, and Jan A. Nowak May 21, 2017 Sepulveda AR, Hamilton SR, Allegra CJ, et al: Molecular Biomarkers for the Evaluation of Colorectal Cancer: Guideline From the American Society for Clinical Pathology, College of American Pathologists, Association for Molecular Pathology, and the American Society of Clinical Oncology. Journal of Clinical Oncology 35:1453-1486, 2017 https://ascopubs.org/doi/full/10.1200/JCO.2016.71.9807
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.	"Colorectal carcinoma patients being considered for anti-EGFR therapy must receive RAS mutational testing. Mutational analysis should include KRAS and NRAS codons 12, 13 of exon 2; 59, 61 of exon 3; and 117 and 146 of exon 4 ("expanded" or "extended" RAS)"
Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	Strength of Evidence: convincing/adequate, benefits outweigh harms; Quality of Evidence: high/intermediate • Convincing: High confidence that available evidence reflects true effect. Further research is very unlikely to change the confidence in the estimate of effect. High/intermediate quality of evidence. • Adequate: Moderate confidence that available evidence reflects true effect. Further research is likely to have an important impact on the confidence in estimate of effect and may change

	the estimate. Intermediate/low quality of evidence.
Provide all other grades and definitions from the evidence grading system	Grades for Strength of Evidence • Convincing: High confidence that available evidence reflects true effect. Further research is very unlikely to change the confidence in the estimate of effect. High/intermediate quality of evidence. • Adequate: Moderate confidence that available evidence reflects true effect. Further research is likely to have an important impact on the confidence in estimate of effect and may change the estimate. Intermediate/low quality of evidence. • Inadequate: Little confidence that available evidence reflects true effect. Further research is very likely to have an important impact on the confidence in the estimate of effect and is likely to change the estimate. Low/insufficient quality of evidence and Expert Panel uses formal consensus process to reach recommendation. • Insufficient: Evidence is insufficient to discern net effect. Any estimate of effect is very uncertain. Insufficient evidence and Expert Panel uses formal consensus process to reach recommendation.
Grade assigned to the **recommendation** with definition of the grade	Recommendation: Some limitations in strength of evidence (adequate or inadequate) and quality of evidence (intermediate or low), balance of benefits and harms, values, or costs, but panel concludes that there is sufficient evidence and/or benefit to inform a recommendation
Provide all other grades and definitions from the recommendation grading system	Grades for Strength of Recommendation • Strong recommendation: Supported by convincing or adequate strength of evidence, high or intermediate quality of evidence, and clear benefit that outweighs any harms. • Recommendation: Some limitations in strength of evidence (adequate or inadequate) and quality of evidence (intermediate or low), balance of benefits and harms, values, or costs,

	but panel concludes that there is sufficient evidence and/or benefit to inform a recommendation. • Expert consensus opinion: Serious limitations in strength of evidence (inadequate of insufficient), quality of evidence (intermediate or low), balance of benefits and harms, values, or costs, but panel consensus is that a statement is necessary. • No recommendation: Insufficient evidence or agreement of the balance of benefits and harms, values, or costs to provide a recommendation.
Body of evidence: • Quantity – how many studies? • Quality – what type of studies?	This recommendation is based on 34 studies, including 29 systematic studies, two meta-analyses, one randomized controlled trial, one prospective cohort study and one retrospective cohort study. Of the 29 systematic reviews, 13 examined important patient subtypes and one reported accounting for patient preferences. Only 3 of the systematic reviews reported using a multidisciplinary panel. All but one included robust method sections. Nine of the systematic reviews rated the quality and strength of evidence reported. The systematic reviews were assessed by the authors for risk of bias: two systematic reviews were deemed low risk, 14 low to moderate risk, 12 moderate risk of bias and one was deemed to have a high risk of bias. Of the two meta-analyses, both included robust and reproducible methods sections, described the planned pooling a priori and discussed limitations of their analysis. Neither meta-analyses provided an assessment of the quality of the studies included. The systematic reviews were assessed by the authors for risk of bias: one was deemed to have a low to moderate risk of bias and the other a moderate risk of bias. The single RCT did not report on the details of randomization but did report on differences in baseline patient characteristics. The RTC was deemed to have a low to moderate risk of bias. The single prospective cohort study reported a balance between the treatment and assessment groups,

	reported baseline patient characteristics, and made adjustments in the analysis accordingly. The prospective cohort study was deemed to have a low risk of bias. The single retrospective cohort study reported balance between the treatment and assessment groups and reported baseline patient characteristics but did not make adjustments in the analysis to account for differences where found. The retrospective cohort study was deemed to have a low risk of bias. All of the evidence that supported this recommendation was assessed and no methodologic flaws were found to raise concerns regarding the findings.
Estimates of benefit and consistency across studies	The evidence described in the studies are directly relevant to this measure, as these data support knowing a patient's tumor mutation status before consideration of use of an EGFR inhibitor in the treatment regimen. Mutational status provides clinically actionable information as negative predictors of benefit to anti-EGFR monoclonal antibody therapies for targeted therapy of colorectal cancer. Early studies included only mutations of KRAS exon 2; however, a large body of evidence is now available to support current guideline recommendations for expanded RAS testing. The evidence is consistent in showing that in addition to mutations in KRAS exon 2, additional RAS mutations in KRAS exons 3 and 4 and NRAS exons 2, 3 and 4 are associated with nonresponse of metastatic colorectal cancer to anti-EGFR monoclonal antibody therapy. The studies are consistent in indicating that EGFR inhibitors (cetuximab and panitumumab) should only be prescribed for patients with metastatic colorectal cancer that are nonmutated/wild type for all known RAS-activating mutations. A reanalysis of the Panitumumab Randomized control Trial in Combination with Chemotherapy for Metastatic Colorectal Cancer to Determine Efficacy (PRIME) trial reported that patients with any RAS mutations were associated with inferior progression free survival (PFS) and overall survival (OS) with anti-EGFR treatment. These findings are consistent with previously reported findings for patients with KRAS mutations in exon 2. A meta-analysis of nine RCT's subsequently provided further evidence that not all KRAS exon 2

	nonmutated/wild-type tumors benefit from anti-EGFR monoclonal antibody treatment in metastatic colorectal cancer; patients with colorectal cancer that are KRAS exon 2 nonmutated/wildtype but harbor RAS mutations in KRAS exons 3 and 4 or NRAS exons 2, 3 and 4 also have significantly inferior anti-EGFR treatment outcomes benefit compared to patients without any RAS mutations. The data show that the clinical benefit of using EGFR inhibitors in treating metastatic colorectal cancer, either as monotherapy or in combination with other treatment regimens, is not seen in patients with RAS-mutated tumors. These data support knowing a patient's tumor mutation status before consideration of use of an EGFR inhibitor in the treatment regimen. Identifying patients whose tumors express mutated RAS will avoid exposing patients to ineffective drugs, avoid exposure to unnecessary drug toxicities, and expedite the use of the best available alternative therapy.
What harms were identified?	The benefits of undergoing testing to determine RAS status outweigh the potential harms associated with a therapy that does not have any efficacy because of RAS status.
Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	Updated guidelines continue to support this measure. Early studies included only mutations of KRAS exon 2; however, a large body of evidence is now available to support current guideline recommendations for expanded RAS testing and this measure has been maintained accordingly.

Source of Systematic Review: • **Title** • **Author** • **Date** • **Citation, including page number** • **URL**	NCCN Guidelines Version 3.2019 Colon Cancer National Comprehensive Cancer Network Version 3.2019 – September 26, 2019 NCCN Clinical Practice Guidelines in Oncology™. Colon Cancer, V.3.2019 (MS-30) https://www.nccn.org  NCCN colon guideline.pdf

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.	"All patients with metastatic colorectal cancer should have tumor tissue genotyped for RAS (KRAS and NRAS) and BRAF mutations individually or as part of a next-generation sequencing (NGS) panel. Patients with any known KRAS mutation (exon 2, 3, 4) or NRAS mutation (exon 2, 3, 4) should not be treated with either cetuximab or panitumumab." "A sizeable body of literature has shown that tumors with a mutation in codon 12 or 13 of exon 2 of the KRAS gene are essentially insensitive to cetuximab or panitumumab therapy... More recent evidence shows mutations in KRAS outside of exon 2 and mutations in NRAS are also predictive for a lack of benefit of anti-EGFR therapies. The panel therefore strongly recommends RAS (KRAS/NRAS) genotyping of tumor tissue (either primary tumor or metastasis) in all patients with metastatic colorectal cancer. Patients with known KRAS or NRAS mutations should not be treated with either cetuximab or panitumumab, either alone or in combination with other anticancer agents, because they have virtually no chance of benefit and the exposure to toxicity and expense cannot be justified. It is implied throughout the guidelines that NCCN recommendations involving cetuximab or panitumumab relate only to patients with disease characterized by RAS wild-type genes. ASCO released a Provisional Clinical Opinion Update on extended RAS testing in patients with metastatic colorectal cancer that is consistent with the NCCN Panel's recommendations. A guideline on molecular biomarkers for colorectal cancer developed by the ASCP, CAP, AMP and ASCO also recommends RAS testing consistent with the NCCN recommendations" (MS-43)
Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	Category 2A: Based upon lower-level evidence, there is uniform NCCN consensus that the intervention is appropriate.
Provide all other grades and definitions from the evidence grading system	NCCN Categories of Evidence and Consensus: • Category 1: Based upon high-level evidence, there is uniform NCCN consensus that the intervention is appropriate.

	• Category 2A: Based upon lower-level evidence, there is uniform NCCN consensus that the intervention is appropriate. • Category 2B: Based upon lower-level evidence, there is NCCN consensus that the intervention is appropriate. • Category 3: Based upon any level of evidence, there is major NCCN disagreement that the intervention is appropriate.
Grade assigned to the **recommendation** with definition of the grade	Category 2A: Based upon lower-level evidence, there is uniform NCCN consensus that the intervention is appropriate.
Provide all other grades and definitions from the recommendation grading system	NCCN Categories of Evidence and Consensus: • Category 1: Based upon high-level evidence, there is uniform NCCN consensus that the intervention is appropriate. • Category 2A: Based upon lower-level evidence, there is uniform NCCN consensus that the intervention is appropriate. • Category 2B: Based upon lower-level evidence, there is NCCN consensus that the intervention is appropriate. • Category 3: Based upon any level of evidence, there is major NCCN disagreement that the intervention is appropriate.
Body of evidence: • Quantity – how many studies? • Quality – what type of studies?	The NCCN guidelines does not include an overview of the body of evidence used for the recommendations specific to RAS mutation status. However, the guidelines does provide an in-depth discussion on the evidence, benefits and harms of EGFR inhibitors. The NCCN guideline presents this data for KRAS Exon 2 Mutations separately from NRAS and other KRAS mutations. This analysis includes the following summary (MS-44 to MS-45): • "KRAS Exon 2 Mutations: A sizeable body of literature has shown that these KRAS exon 2 mutations are predictive of lack of response to cetuximab or panitumumab therapy, and FDA labels for cetuximab and panitumumab specifically state that these agents are not recommended for the treatment of colorectal cancer characterized by these mutations. Results

	are mixed as far as the prognostic value of KRAS mutations. In the Alliance N0147 trial, patients with KRAS exon 2 mutations experienced a shorter DFS than patients without such mutations. At this time, however, the test is not recommended for prognostic reasons. A retrospective study from De Roock et al raised the possibility that codon 13 mutations (G13D) in KRAS may not be absolutely predictive of non-response. Another retrospective study showed similar results. However, more recent retrospective analysis of 3 randomized controlled phase III trials concluded that patients with KRAS G13D mutations were unlikely to respond to panitumumab. Results from a prospective phase II single-arm trial assessed the benefit of cetuximab monotherapy in 12 patients with refractory metastatic colorectal cancer whose tumors contained KRAS G13D mutations. The primary endpoint of 4-month progression-free rate was note met (25%), AND NO RESPONSES WERE SEEN. Preliminary results of the AGITG phase II ICE CREAM trial also failed to see a benefit of cetuximab monotherapy in patients with KRAS G13D mutations. However, partial responses were reported after treatment with irinotecan plus cetuximab in 9% of this irinotecan-refractory population. A meta-analysis of 8 RCTs came to the same conclusion: that tumors with KRAS G13D mutations are no more likely to respond to EGFR inhibitors than tumors with other KRAS mutations. The panel believes that patients with any known KRAS mutation, including G13D, should not be treated with cetuximab or panitumumab. (MS-44) • "NRAS and Other KRAS Mutations: In the AGITG MAX study, 10% of patients with wild-type KRAS exon 2 had mutations in KRAS exons 3 or 4 or in NRAS exons 2, 3 and 4. In the PRIME trial, 17% of 641 patients without KRAS exon 2 mutations were found to have mutations in exons 3 and 4 of KRAS or mutations in exons 2, 3 and 4 or NRAS. A predefined retrospective subset analysis of data from PRIME revealed that PFS (HR, 1.31; 95% CI, 1.07-1.60; P = .008) and OS (HR, 1.21; 95% CI, 1.01-1.45; P= .04) were decreased in patients with any KRAS or NRAS mutation who received panitumumab plus FOLFOX compared to those who received FOLFOX alone. These results show that panitumumab does not benefit

	patients with KRAS or NRAS mutations and may even have a detrimental effect in these patients. Updated analysis of the FIRE-3 trial… was recently published. When all RAS (KRAS/NRAS) mutations were considered, PFS was significantly worse in patients with RAS-mutant tumors receiving FOLFIRI plus cetuximab than in patients with RAS-mutant tumors receiving FOLFIRI plus bevacizumab (6.1 months vs. 12.2 months; P = .004). on the other hand, patients with KRAS/NRAS wild-type tumors showed no difference in PFS between the regimens (10.4 months vs. 10.2 months, P = .54). This result indicates that cetuximab likely has a detrimental effect in patients with KRAS or NRAS mutations The FDA indication for panitumumab was recently updated to state that panitumumab is not indicated for the treatment of patients with KRAS or NRAS mutation-positive disease in combination with oxaliplatin-based chemotherapy. The NCCN Colon/Rectal Cancer Panel believes that RAS mutation status should be determined as diagnosis of stage IV disease. Patients with any known RAS mutation should not be treated with either cetuximab or panitumumab. (MS-44/MS-45)
Estimates of benefit and consistency across studies	See Body of Evidence section.
What harms were identified?	See Body of Evidence section.
Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	Updated guidelines continue to support this measure.

1a.4 OTHER SOURCE OF EVIDENCE

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

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

1a.4.2 What process was used to identify the evidence? 1a.4.3. Provide the citation(s) for the evidence.

1b. Performance Gap

Demonstration of quality problems and opportunity for improvement, i.e., data demonstrating:

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

1b.1. Briefly explain the rationale for this measure *(e.g., how the measure will improve the quality of care, the benefits or improvements in quality envisioned by use of this measure)*

<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.

We envision that use of this measure will improve concordance with recommendations for expanded RAS testing. Evidence now supports testing for NRAS in addition to KRAS mutations. ASCO anticipates a greater performance gap due to the guideline update, which is a relatively new requirement in the field. Clinical trials data show that the benefit of using EGFR inhibitors in treating metastatic colorectal cancer, either as monotherapy or in combination with other treatment regimens, is limited to non-existent in patients with RAS-mutated tumors. These data strongly suggest that patients with RAS mutations are better served with other therapies, especially considering the harms and costs of anti-EGFR treatment.

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.*

Testing to identify statistically significant and meaningful differences in performance was conducted using 2017 MIPS performance from registry data provided from CMS. The 2017 data was from 129 providers representing 41 practices and 375 individual patients. Practices were identified by unique number of TINs and individual clinicians were identified by unique number of NPIs. Additional descriptive characteristics of the measured entities, such as size and location type, are unknown. Entities submitted data for inclusion in this data set according to the eligibility and reporting requirements for MIPS 2017 program year. Measures of central tendency, variability and dispersion were calculated. Measures of central tendency, variability and dispersion were calculated. We were unable to determine from our rolled-up data sample the number of clinicians who reported to MIPS as an individual or group; therefore, this measure should be considered for endorsement at the group/practice level, with a potential group size as n of 1 or group of 1.

Data collected from the 2017 MIPS reporting year demonstrates variation and room for improvement: practice mean = 76.1% with a confidence interval (0.65, 0.87); practice minimum = 0%; practice maximum = 100%; practice percent outside confidence interval = 80.49%. For 2017 MIPS reporting, individual clinician mean = 80.67% with a confidence

interval (0.75, 0.87); individual clinician minimum = 0%; individual clinician maximum = 100%; individual clinician percent outside confidence interval = 99.22%.

Additional details from the TIN-level analysis are provided below.

Number of unique entities:

Frequency

43

Denominators

Min	Q1	Median	Mean	Q3	Max
1	2	6	11.51	12	82

Measure Distribution:

Min	Q1	Median	Mean	Q3	Max	CI for mean	Percent outside CI
0	0.9902	1	0.9123	1	1	(0.85, 0.98)	95.35

An analysis at the TIN level indicated that while a slight majority (approximately 54%) of practices perform at 100% there are meaningful differences in performance across practices. Multiple practices perform at lower levels with the lowest performance score at 0% and average performance of 76% indicating room for improvement in a significant portion of practices. It should be noted that small sample size may impact the results presented, as the median denominator is 3, meaning that half of the performance in the graph above are based 3 patients or less.

It should be noted that performance data from MIPS data 2017 does not include data for expanded RAS testing as those changes were implemented in 2018. We do not believe that the measure has been substantively changed in regard to its impact on reliability and validity as the data fields used and the clinical work flow remain the same; however, we do anticipate a greater performance gap due to the guideline update, which is a relatively new requirement in the field.

Data collected in the Fall 2011 QOPI round demonstrates variation and room for improvement, with a range of 33%-100%, mean 73% (N=151 patient records, 18 practices).

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.

49%-82% of colorectal tumors have reported overexpression in EGFR (1). Anti-EGFR monoclonal antibodies (cetuximab and panitumumab) inhibit the downstream signaling pathways in EGFR but are effective in only 10-20% of patients with colorectal cancer because of mutations in pathways downstream of EGFR, including RAS mutations (1). Earlier studies and guidelines recommendations included only mutations of KRAS exon 2.

A population-based study using data collected by Surveillance, Epidemiology and End Results (SEER) registries found the overall proportion of KRAS testing was only 22.7% among Stage IV patients with substantial variation by geographic region and patient characteristics (2). They identified wide variation in documented KRAS testing for Stage IV colorectal patients, with rates ranging from 15% in Louisiana to 39% in New Mexico (2). Demographic characteristics associated with higher proportions of KRAS testing included a younger age, white or other race, being married and living in an urban area (2).

Similarly, a 2017 population-based study using 2010-2013 data from the New Mexico Tumor Registry reported KRAS testing was completed in 38.4% of patients and identified age and geographic disparities (3).

Newer evidence is now available to support current guideline recommendations for expanded RAS testing to identify RAS mutations in KRAS exons 2, 3 and 4 and NRAS exons 2, 3 and 4; however, data on guideline adherence is limited as the recommendations were released in 2017. In the AGITG MAX study, 10% of patients with wild-type KRAS exon 2 status had another RAS mutation (4). In the PRIME trial, 17% of patients without KRAS exon 2 mutations had another

RAS mutation (5). These populations represent additional opportunity for improvement in the completion of expanded RAS testing for patients with advanced colorectal cancer.

1. NCCN Clinical Practice Guidelines in Oncology™. Colon Cancer, V.3.2019 https://www.nccn.org/professionals/physician_gls/pdf/colon.pdf

2. Charlton, M. E., Karlitz, J. J., Schlichting, J. A., Chen, V. W., & Lynch, C. F. (2017). Factors Associated With Guideline-recommended KRAS Testing in Colorectal Cancer Patients: A Population-based Study. American journal of clinical oncology, 40(5), 498–506. doi:10.1097/COC.0000000000000191

3. Greenbaum, A., Wiggins, C., Meisner, A. L., Rojo, M., Kinney, A. Y., & Rajput, A. (2017). KRAS biomarker testing disparities in colorectal cancer patients in New Mexico. Heliyon, 3(11), e00448.

4. Price, T. J., Bruhn, M. A., Lee, C. K., Hardingham, J. E., Townsend, A. R., Mann, K. P., ... & Gebski, V. (2015). Correlation of extended RAS and PIK3CA gene mutation status with outcomes from the phase III AGITG MAX STUDY involving capecitabine alone or in combination with bevacizumab plus or minus mitomycin C in advanced colorectal cancer. British journal of cancer, 112(6), 963.

5. Douillard, J. Y., Oliner, K. S., Siena, S., Tabernero, J., Burkes, R., Barugel, M., ... & Rivera, F. (2013). Panitumumab–FOLFOX4 treatment and RAS mutations in colorectal cancer. New England Journal of Medicine, 369(11), 1023-1034.

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.*

While this measure is included in the MIPS program, this program has not yet made disparities data available for ASCO to analyze the report.

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

A population-based study using data collected by Surveillance, Epidemiology and End Results (SEER) registries found the overall proportion of KRAS testing was only 22.7% among Stage IV patients with substantial variation by geographic region and patient characteristics (1). They identified wide variation in documented KRAS testing for Stage IV colorectal patients, with rates ranging from 15% in Louisiana to 39% in New Mexico (1). Demographic characteristics associated with higher proportions of KRAS testing included a younger age, white or other race, being married and living in an urban area (1). Similarly, a 2017 population-based study using 2010-2013 data from the New Mexico Tumor Registry reported KRAS testing was completed in 38.4% of patients and identified age and geographic disparities (2).

Newer evidence is now available to support current guideline recommendations for expanded RAS testing to identify RAS mutations in KRAS exons 2, 3 and 4 and NRAS exons 2, 3 and 4; however, data on guideline adherence is limited as the recommendations were released in 2017. It is expected that the same geographic and demographic characteristics associated with low concordance to KRAS testing are also associated with concordance to expanded RAS testing guidelines.

1.Charlton, M. E., Karlitz, J. J., Schlichting, J. A., Chen, V. W., & Lynch, C. F. (2017). Factors Associated With Guideline-recommended KRAS Testing in Colorectal Cancer Patients: A Population-based Study. American journal of clinical oncology, 40(5), 498–506. doi:10.1097/COC.0000000000000191

2.Greenbaum, A., Wiggins, C., Meisner, A. L., Rojo, M., Kinney, A. Y., & Rajput, A. (2017). KRAS biomarker testing disparities in colorectal cancer patients in New Mexico. Heliyon, 3(11), e00448.

2. Reliability and Validity—Scientific Acceptability of Measure Properties

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

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

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

Cancer, Cancer : Colorectal

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

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.)*

https://qpp.cms.gov/docs/QPP_quality_measure_specifications/CQM-Measures/2019_Measure_451_MIPSCQM.pdf

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

This is not an eMeasure **Attachment:**

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

No data dictionary **Attachment:**

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

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

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

Not an instrument-based measure

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

Yes

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

This measure has been expanded to RAS mutational testing based on a guideline update to include NRAS in addition to KRAS. In addition to testing for mutations in KRAS exon 2 (codons 12 and 13) as recommended previously, before

treatment with anti-EGFR antibody therapy, patients with metastatic colorectal cancer should have their tumor tested for mutations in:

• KRAS exons 3 (codons 59 and 61) and 4 (codons 117 and 146)

• NRAS exons 2 (codons 12 and 13), 3 (codons 59 and 61), and 4 (codons 117 and 146)

This measure is based on an ASCO Guideline:

"Colorectal carcinoma patients being considered for anti-EGFR therapy must receive RAS mutational testing. Mutational analysis should include KRAS and NRAS codons 12, 13 of exon 2; 59, 61 of exon 3; and 117 and 146 of exon 4 ("expanded" or "extended" RAS)".

Sepulveda AR, Hamilton SR, Allegra CJ, et al: Molecular Biomarkers for the Evaluation of Colorectal Cancer: Guideline From the American Society for Clinical Pathology, College of American Pathologists, Association for Molecular Pathology, and the American Society of Clinical Oncology. Journal of Clinical Oncology 35:1453-1486, 2017

Additionally, we removed exclusion for patient transfer to practice after initiation of chemotherapy. We believe this constitutes a non-substantive change.

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

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

RAS (KRAS and NRAS) gene mutation testing performed prior to initiation of anti-EGFR monoclonal antibody therapy

S.5. Numerator Details *(All information required to identify and calculate the cases from the target population with the target process, condition, event, or outcome such as definitions, time period for data collection, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*

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

RAS gene mutation testing = RAS mutation detected

OR

RAS gene mutation testing = No RAS mutation detected (wildtype)

AND

RAS gene mutation testing date

Numerator definitions:

RAS mutation testing - RAS testing for this measure refers to assays that detect mutations in codons 12 and 13 of exon 2, codons 59 and 61 of exon 3 and codons 117 and 146 in exon 4 in KRAS or NRAS. Do not include results from mutations at other codons or assays for other alterations (e.g., BRAF, PI3K, PTEN genes). The College of American Pathologists (CAP) Perspectives on Emerging Technology (POET) Report on RAS mutation testing provides additional guidance on testing.

If multiple RAS mutation tests have been performed, refer to the most recent test results.

In the absence of any documentation regarding testing for the RAS gene mutation, select 'Test not ordered/no documentation.'

Refer to the interpretive report for the RAS test. The report will indicate if a mutation within codons 12 and 13 of exon 2, codons 59 and 61 of exon 3 and codons 117 and 146 in exon 4 in KRAS or NRAS, where KRAS or NRAS gene was detected in the DNA extracted from the colon tumor specimen.

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

Adult patients with metastatic colorectal cancer who receive anti-EGFR monoclonal antibody therapy

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

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

Age at diagnosis greater than or equal to 18 years

AND

2 or more encounters at the reporting site

AND

Initial colon or rectal cancer diagnosis (153.x, 154.0, 154.0, 154.1, 154.8)

AND

Presence of metastatic disease documented

AND

Anti-EGFR monoclonal antibody therapy received

Definitions

Encounter: new patient visit (CPT 99201-99205) or established patient (CPT 99211-99215), not consult (CPT 99241-99245) office consult or inpatient consult CPT 99251-99255)

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

None

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.)*

n/a

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

n/a

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

No risk adjustment or risk stratification

If other:

S.12. Type of score:

Rate/proportion

If other:

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

Better quality = Higher score

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

This measure is a proportion without exclusions. The calculation algorithm is: (Patients meeting the numerator/patients in the denominator) x 100

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.

Measure is not based on a sample.

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.

Paper Medical Records, 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.

N/A, measure is not instrument-based.

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 : Group/Practice

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.)*

2. Validity – See attached Measure Testing Submission Form

1859_MeasureTesting_MSF5.0_Data.doc,1859_nqf_testing_attachment_073019_FINAL-637001802906265569.docx

2.1 For maintenance of endorsement

Reliability testing: If testing of reliability of the measure score was not presented in prior submission(s), has reliability testing of the measure score been conducted? If yes, please provide results in the Testing attachment. Please use the

most current version of the testing attachment (v7.1). Include information on all testing conducted (prior testing as well as any new testing); use red font to indicate updated testing.

Yes

2.2 For maintenance of endorsement

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

Yes

2.3 For maintenance of endorsement

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

No - This measure is not risk-adjusted

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

Measure Number (*if previously endorsed*)**:** 1860

Measure Title: Patients with metastatic colorectal cancer and RAS gene mutation spared treatment with anti-epidermal growth factor receptor monoclonal antibodies

Date of Submission: TBD

Type of Measure:

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

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

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

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

Measure Specified to Use Data From: (***must be consistent with data sources entered in S.17***)	**Measure Tested with Data From:**
☒abstracted from paper record	☒abstracted from paper record
☐ claims	☐ claims

☒registry	☒registry
☐ abstracted from electronic health record	☐ abstracted from electronic health record
☐ eMeasure (HQMF) implemented in EHRs	☐ eMeasure (HQMF) implemented in EHRs
☐ other: Click here to describe	☐ other: Click here to describe

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

The datasets used for testing were 2011 QOPI data and 2017 MIPS data, which are consistent with the measure specifications.

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

Data reported from QOPI are from the fall 2011 QOPI round, reflecting data submitted October and November 2011. Data reported from MIPS are from 2017. The MIPS performance year begins on January 1 and ends December 31 each year. MIPS program participants must report data collected during one calendar year by March 31 of the following calendar year.

1.4. What levels of analysis were tested? (*testing must be provided for <u>all</u> 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 <u>measured entities</u> 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*)

<u>2019 Submission:</u>

Testing to identify statistically significant and meaningful differences in performance was conducted using 2017 MIPS performance from registry data provided from CMS. The 2017 data was from 158 providers representing 43 practices and 495 individual patients. Practices were identified by unique number of TINs and individual clinicians were identified by unique number of NPIs. Additional descriptive characteristics of the measured entities, such as size and location type, are unknown. Entities submitted data for inclusion in this data set according to the eligibility and reporting requirements for MIPS 2017 program year. Measures of central tendency, variability and dispersion were calculated. Measures of central tendency, variability and dispersion were calculated. We were unable to determine from our rolled-up data sample the number of clinicians who reported to MIPS as an individual or group; therefore, this measure should be considered for endorsement at the group/practice level, with a potential group size as n of 1 or group of 1.

2012 Submission:

Data reported are from the Fall 2011 QOPI round, reflecting data submitted October and November 2011. 136 practices reported this measure. Data from 444 patient records were submitted for this measure.

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*)

2019 Submission:

Data from a total of 495 patient records were submitted for this measure.

2012 Submission:

QOPI measure analytics at the practice level were generated. Data from 444 patient records were submitted 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.

2019 Submission:

Testing data was supplemented with 2017 MIPS performance data to identify statistically significant and meaningful differences in performance.

2012 Submission:

Testing data are from the fall 2011 QOPI round (reflecting data submitted October and November 2011). The QOPI data was used to perform data element validity testing in the 2012 submission.

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 points for social risk factors were not available to perform an analysis.

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*)

2019 Submission:

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 (facility-to-facility) / [Variance (facility-to-facility) + Variance (facility-specific-error]

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

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

To assess signal-to-noise, we employed the beta-binomial model as described by JL Adams (1). Each facility provided numerators and denominators in accordance with the measure specification. Through the estimation of the beta-binomial parameters (often referred to as alpha and beta) as described by Adams (1), we estimated the facility-to-facility variance and the within-facility variance (simply the binomial variance for each facility).

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 practice 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.)

2012 Submission:

Data/Sample 2010-2011 audit: QOPI practices applying for the QOPI Certification Program are required to submit copies of documentation from 3-5 records which were previously abstracted. Trained ASCO auditors randomly select records within each domain for audit. Agreement at the data element level is documented. 426 audited records from 130 practices were complete in November 2011 and included in the concordance analysis.

Analytic method 2010-2011 audit: Agreement data from 426 records were imported into a formatted data table for analysis. First, agreement data were used to calculate concordance at the data element level. Second, by applying the measure analytic calculation, concordance at the measure level was calculated.

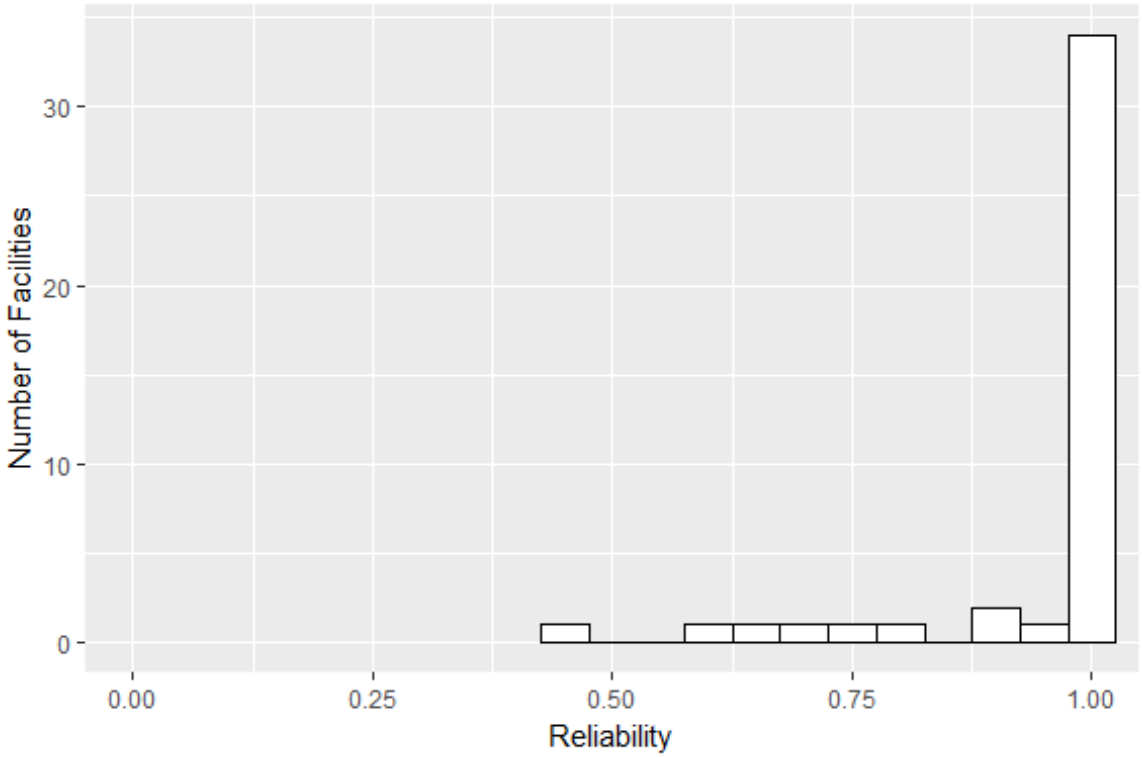
2a2.3. For each level of testing checked above, what were the statistical results from reliability testing? (e.*g., percent agreement and kappa for the critical data elements; distribution of reliability statistics from a signal-to-noise analysis*)

2019 Submission:

Signal to noise analysis using the Beta-Binomial determined mean reliability is 89%, with a median of 100%.

Facility-level Reliability

N	Alpha	Beta	Min	10th Pctl	Median	90th Pctl	Max	Mean
43	0.8322	0.1187	0.4305	0.7439	1	1	1	0.9465



2012 Submission:

Testing results 2010-2011 audit: measure level concordance 90% (valid N=145 records)

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?*)

2019 Submission:

Signal to noise analysis using the Beta-Binomial yielded a reliability greater than 0.90, which is considered very high. The mean reliability of 95% observed is categorized as high reliability and the 10th percentile is 74%; thus, thus, reliability is acceptable.

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)*

2019 Submission:

Correlation analysis was completed to conduct empirical validity testing using 2017 MIPS data. KRAS gene mutation testing performed for patients with metastatic colorectal cancer who receive anti-epidermal growth factor receptor monoclonal antibody therapy (QI #451/ NQF#1859) 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 with metastatic colorectal cancer and KRAS gene mutation spared treatment with anti-epidermal growth factor receptor monoclonal antibodies (NQF #1860) and patients with metastatic colorectal cancer who receive anti-epidermal growth factor receptor monoclonal antibody therapy for whom KRAS gene mutation testing was performed (NQF #1859).

Datasets were reviewed to identify shared providers based on TIN identifiers. Correlation analysis was performed to evaluate the association between performance scores of these shared practices.

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.

2012 Submission:

In 2009, an ASCO steering group comprised of medical oncologists, health services researchers, and quality experts undertook an iterative, criteria-based assessment process to identify QOPI measures that are appropriate for use for accountability measurement. This measure was selected as appropriate for accountability.

Face validity of the measure score was assessed via survey of experts involved in ASCO committees in 2011. The survey explicitly asked whether the scores obtained from the measure as specified will provide an accurate reflection of quality and can be used to distinguish good and poor quality.

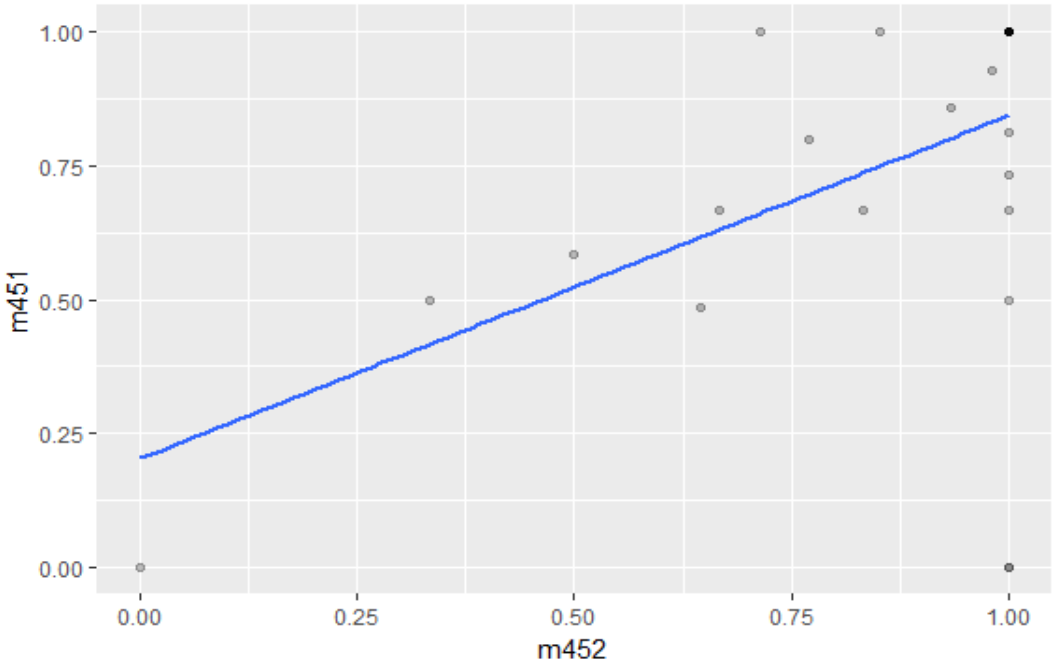
2b1.3. What were the statistical results from validity testing? (*e.g., correlation; t-test*)

2019 Submission:

Correlation analysis determined patients with metastatic colorectal cancer and KRAS gene mutation spared treatment with anti-epidermal growth factor receptor monoclonal antibodies (QI #452/NQF #1860) is positively

correlated with KRAS gene mutation testing performed for patients with metastatic colorectal cancer who receive anti-epidermal growth factor receptor monoclonal antibody therapy (QI #451/NQF #1859).

The correlation coefficient observed was 0.49 based on 28 matching practices.



2012 Submission:

Face validity survey results revealed that 82% of respondents 'strongly agree' or 'agree' that this measure provides an accurate reflection of quality and can be used to distinguish good and poor quality.

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?*)

2019 Submission

This measure has a strong positive correlation with another evidence-based process of care, as the correlation coefficient observed of 0.49 is greater than the 0.40 threshold for interpretation of a strong correlation. The correlation demonstrates the criterion validity of the measure

2012 Submission:

Face validity testing demonstrated a majority of respondents (82%) strongly agree or agree that the measure provided an accurate reflection of quality and can be used to distinguish good and poor quality.

2b2. EXCLUSIONS ANALYSIS

NA ☒ no exclusions — *skip to section 2b3*

2b2.1. Describe the method of testing exclusions and what it tests (*describe the steps—do not just name a method; what was tested, e.g., whether exclusions affect overall performance scores; what statistical analysis was used*)

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

2b2.3. What is your interpretation of the results in terms of demonstrating that exclusions are needed to prevent unfair distortion of performance results? (*i.e., the value outweighs the burden of increased data collection and analysis. Note: **If patient preference is an exclusion**, the measure must be specified so that the effect on the performance score is transparent, e.g., scores with and without exclusion*)

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

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

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

- ☐ **No risk adjustment or stratification**
- ☐ **Statistical risk model with** Click here to enter number of factors **risk factors**
- ☐ **Stratification by** Click here to enter number of categories **risk categories**
- ☐ **Other,** Click here to enter description

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

2b3.2. If an outcome or resource use component measure is not risk adjusted or stratified, provide rationale and analyses to demonstrate that controlling for differences in patient characteristics (case mix) is not needed to achieve fair comparisons across measured entities.

2b3.3a. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors (clinical factors or social risk factors) used in the statistical risk model or for stratification by risk (*e.g., potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of p<0.10; correlation of x or higher; patient factors should be present at the start of care*)

Also discuss any "ordering" of risk factor inclusion; for example, are social risk factors added after all clinical factors?

2b3.3b. How was the conceptual model of how social risk impacts this outcome developed? Please check all that apply:

- ☐ **Published literature**
- ☐ **Internal data analysis**
- ☐ **Other (please describe)**

2b3.4a. What were the statistical results of the analyses used to select risk factors?

2b3.4b. Describe the analyses and interpretation resulting in the decision to select social risk factors *(e.g. prevalence of the factor across measured entities, empirical association with the outcome, contribution of unique variation in the outcome, assessment of between-unit effects and within-unit effects.)* **Also describe the impact of adjusting for social risk (or not) on providers at high or low extremes of risk.**

2b3.5. Describe the method of testing/analysis used to develop and validate the adequacy of the statistical model or stratification approach (*describe the steps—do not just name a method; what statistical analysis was used*)

Provide the statistical results from testing the approach to controlling for differences in patient characteristics (case mix) below.

If stratified, skip to 2b3.9

2b3.6. Statistical Risk Model Discrimination Statistics (*e.g., c-statistic, R-squared*)**: 2b3.7. Statistical Risk Model Calibration Statistics** (*e.g., Hosmer-Lemeshow statistic*): **2b3.8. Statistical Risk Model Calibration – Risk decile plots or calibration curves**: **2b3.9. Results of Risk Stratification Analysis**:

2b3.10. What is your interpretation of the results in terms of demonstrating adequacy of controlling for differences in patient characteristics (case mix)? (i.*e., what do the results mean and what are the norms for the test conducted*)

2b3.11. Optional Additional Testing for Risk Adjustment (*not required, but would provide additional support of adequacy of risk model, e.g., testing of risk model in another data set; sensitivity analysis for missing data; other methods that were assessed*)

2b4. IDENTIFICATION OF STATISTICALLY SIGNIFICANT & MEANINGFUL DIFFERENCES IN PERFORMANCE

2b4.1. Describe the method for determining if statistically significant and clinically/practically meaningful differences in performance measure scores among the measured entities can be identified

(*describe the steps—do not just name a method; what statistical analysis was used? Do not just repeat the information provided related to performance gap in 1b)*

2019 Submission:

We defined a meaningful difference as the presence of a significant spread between the minimum and maximum scores or a significant spread between median and either the minimum or maximum scores. A significant spread between the 25th and 75th percentile (the inner-quartile range [IQR]) was also considered to represent a meaningful difference. Therefore, we calculated several descriptive statistics, including the minimum, maximum, 25th and 75th percentile, median, IQR, and range. Additionally, we calculated the standard deviation, standard error of the mean performance, and 95% confidence interval for the mean performance. Finally, we calculated the percent of facilities whose performance was statistically significantly different from the overall performance mean

2012 Submission:

QOPI measure analytics at the practice level were generated.

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*)

2019 Submission:

For 2017 MIPS reporting, practice mean = 91.23% with a confidence interval (0.85, 0.98); practice minimum = 0%; practice maximum = 100%; practice percent outside confidence interval = 93.35%. For 2017 MIPS reporting, individual clinician mean = 91.7% with a confidence interval (0.88, 0.95); individual clinician minimum = 0%; individual clinician maximum = 100%; individual clinician percent outside confidence interval = 100%.

Additional details from the TIN-level analysis are provided below.

Number of unique entities

Frequency
41

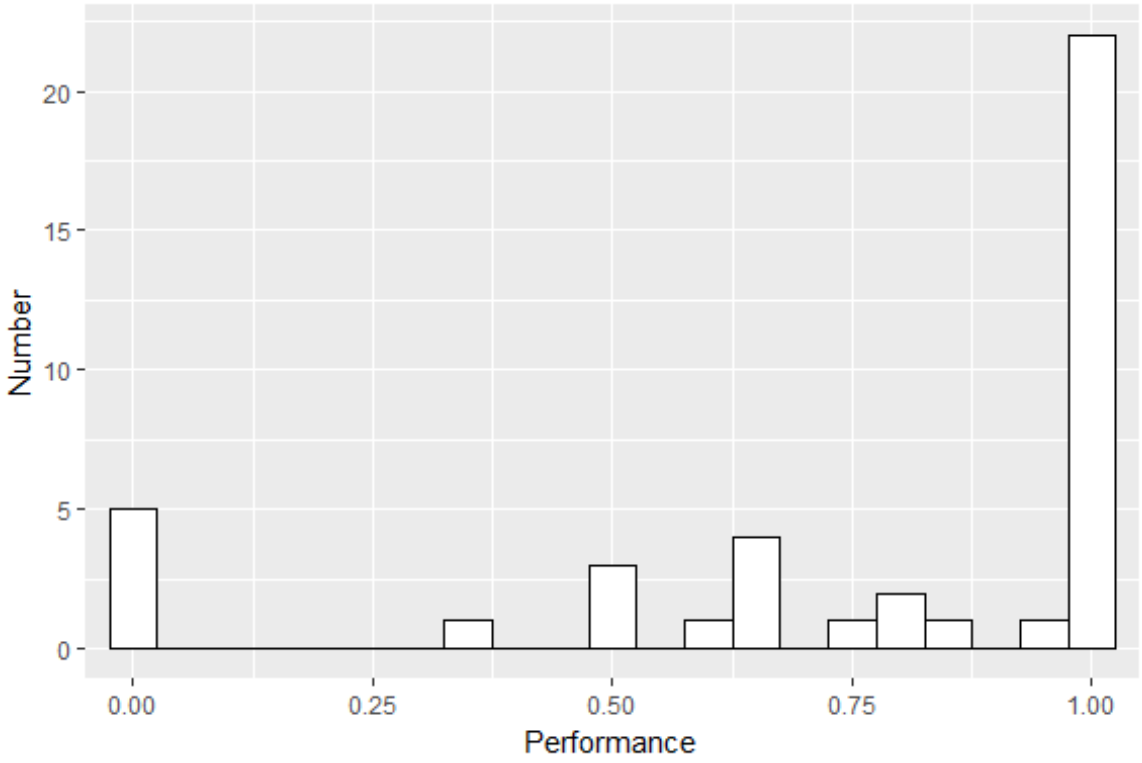
Denominators

Min	Q1	Median	Mean	Q3	Max
1	2	3	9.146	8	101

Measure Distribution

Min	Q1	Median	Mean	Q3	Max	CI for mean	Percent outside CI
0	0.6667	1	0.761	1	1	(0.65, 0.87)	80.49

Measure Distribution:



2012 Submission:

For Fall 2011 QOPI round, practice mean = 85%; practice minimum = 0%; practice maximum = 100%.

*If analytics are limited to practices reporting 5 or more records for this measure, the minimum is 40% and maximum is 100%.

This measure has been implemented in QOPI for several years. In this self-selected group of oncology practitioners committed to quality assessment and improvement, this measure demonstrates sub-optimal variation.

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?*)

2019 Submission:

An analysis at the TIN level indicated a majority (approximately 76.7%) of practices perform at 100% with a mean performance of 91%. The mean performance rate of 91% is statistically significant from 100%, suggesting room for improvement remains across practices.

Performance data from MIPS data 2017 does not include data for expanded RAS testing as those changes were implemented in 2018. We do not believe that the measure has been substantively changed in regard to its impact on reliability and validity as the data fields used and the clinical work flow remain same; however, we do anticipate a greater performance gap due to the guideline update, which is a relatively new requirement in the field.

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

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

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

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

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

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

2b6. MISSING DATA ANALYSIS AND MINIMIZING BIAS

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

The MIPS dataset provided to us from the 2015-2017 program years did not contain missing data, so this test was not performed. Due to data completeness requirements, we suspect that missing data would have been rejected when submitted to CMS, in which case those values would not be counted towards measure performance. While data that may have been missing prior to a submission to CMS is unknown and therefore precluded any analysis, there is no indication that this missing data was systematic, thus their omission would lead to unbiased performance results.

In the QOPI dataset, patients are only included in the denominator if they meet the specified data elements and definitions and practices cannot submit a patient file without completing all of the required data elements for the measure. In addition, the lack of documentation in the medical record that the patient met the numerator requirements would be interpreted as a quality failure. As a result, concerns over missing data are minimized through these data entry requirements and the overall high rate of concordance demonstrated in our data element validity 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*)

2019 Submission:

This test was not performed for this measure as 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*)

2019 Submission:

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

3. Feasibility

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

3a. Byproduct of Care Processes

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

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

Generated or collected by and used by healthcare personnel during the provision of care (e.g., blood pressure, lab value, diagnosis, depression score)

If other:

3b. Electronic Sources

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

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

Some data elements are in defined fields in electronic sources

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

All data elements needed for this measure are collected through electronic data or using keyword searches. ASCO is in the process of assessing the feasibility of developing an electronic clinical quality measure.

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.

Apart from the lack of availability of disparities data for analysis, we have not identified any areas of concern or made any modifications as a result of testing and operational use of this measure in relation to data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, time and cost of data collection, or other feasibility issues unless otherwise noted.

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)*.

ASCO requests interested parties seek a licensing agreement prior to commercial use of this measure.

4. Usability and Use

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

4a. Accountability and Transparency

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

4.1. Current and Planned Use

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

Specific Plan for Use	Current Use (for current use provide URL)
Public Reporting	Payment Program Merit-Based Incentive Payment System (MIPS) Program https://qpp.cms.gov/mips/quality-measures ASCO Qualified Clinical Data Registry https://practice.asco.org/sites/default/files/drupalfiles/QCDR-2019-Measure-Summary.pdf Professional Certification or Recognition Program QOPI® Certification Program https://practice.asco.org/quality-improvement/quality-programs/quality-oncology-practice-initiative/qopi-related-measures Quality Improvement (external benchmarking to organizations) Quality Oncology Practice Initiative (QOPI®) https://practice.asco.org/quality-improvement/quality-programs/quality-oncology-practice-initiative/qopi-related-measures Quality Improvement (Internal to the specific organization) Quality Oncology Practice Initiative (QOPI®) https://practice.asco.org/quality-improvement/quality-programs/quality-oncology-practice-initiative/qopi-related-measures

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

Merit-based Incentive Payment System (MIPS) reporting program, Center for Medicare and Medicaid Services Prior to 2016, this measure was used for Eligible Providers (EPs) in the Physician Quality Reporting System (PQRS). As of 2017, MIPS replaced the PQRS program. MIPS is a national performance-based payment program that uses performance scores across several categories to determine payment rates for EPs. MIPS takes a comprehensive approach to payment by basing consideration of quality on a set of evidence-based measures that were primarily developed by clinicians, thus encouraging improvement in clinical practice and supporting advances in technology that allow for easy exchange of information. Data on geographic area and number and

percentage of accountable entities and patients, including level of measurement and setting, are unavailable for analysis.

Quality Oncology Practice Initiative (QOPI®)

In 2002, the American Society of Clinical Oncology established the Quality Oncology Practice Initiative (QOPI®). QOPI® is an oncologist-led, practice-based quality assessment and improvement program designed to promote excellence in cancer care by helping practices create a culture of self-examination and improvement. QOPI provides a standard methodology, robust library of quality metrics for oncology, and a collection tool to reliably and routinely assess care, inform quality improvement activities, and demonstrate quality to patients and external stakeholders. Collection rounds are offered twice per year, in spring and fall, for an eight-week period. QOPI® continues to be a successful program in the United States and 7 other countries, with 265, 213, 257 and 209 unique practices participating in Round 2 2017, Round 1 2018, Round 2 2018 and Round 1 2019 respectively.

QOPI® Qualified Clinical Data Registry

In addition to the current use for quality improvement with benchmarking in the QOPI® registry, this measure has been reported to CMS by the registry as a Qualified Clinical Data Registry. QOPI® was deemed as a registry for oncology measures group reporting and as a QCDR to report to PQRS in 2015 and 2016 and to report to MIPS in 2017, 2018 and 2019. Eligible professionals will be considered to have satisfactorily participated in MIPS if they submit quality measures data or results to CMS via a qualified clinical data registry. In 2017 and 2018, a total of 19 practices representing approximately 50,000 patient charts submitted to MIPS through QOPI. CMS has implemented a phased approach to public reporting performance information on the Physician Compare website.

QOPI® Certification Program

The QOPI® Certification Program provides a three-year certification for outpatient hematology-oncology practices. To obtain Certification, a practice must achieve an aggregate score above 75% adherence on 26 measures that count toward the overall Quality Score. Please see a description of the QOPI® program above for details.

Core Quality Measure Collaborative's (CQMC) Medical Oncology Core Measure Set

This measure has also been included in the Core Quality Measure Collaborative's (CQMC) Medical Oncology Core Measure Set. The CQMC is a broad-based coalition of health care leaders convened by America's Health Insurance Plans (AHIP) starting in 2015. The purpose of this program is to reduce variability in measure selection, specifications and implementation. The CQMC defines a core measure set as a parsimonious group of scientifically sounds measures that efficiently promote a patient-centered assessment of quality and should be prioritized for adoption in value-based purchasing and APMs. The CQMC has developed and released core sets of quality measures that could be implemented across both commercial and government payers. The measures have been implemented nationally by private health plans using a phased-in approach. Contracts between physicians and private payers are individually negotiated and therefore come up for renewal at different points in time depending on the duration of the contract. It is anticipated that private payers will implement these core sets of measures as and when contracts come up for renewal or if existing contracts allow modification of the performance measure set. CMS is also working to align measures across public program.

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 used for multiple accountability applications and public reporting is forthcoming. According to the CY 2019 Quality Payment Program final rule, Physician Compare has continued to pursue a phased approach to public reporting under MACRA. CMS intends to make all measures under MIPS quality performance category available for public reporting on Physician Compare. These measures include those reported via all available submission methods for MIPS-eligible clinicians and groups. Because this measure has

been in use for at least one year and meets the minimum sample size requirement for reliability, this measure meets criteria for public reporting but has not yet been included in Physician Compare.
As described above, CMS is also planning to publicly report QCDR data. Additionally, although the measure is currently in use, we will continue to seek opportunities to advocate for expanded use of this measure in government or other programs.

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

Despite not yet being included in Physician Compare, this measure meets criteria for public reporting because it has been in use for at least one year and meets the minimum sample size requirement for reliability, this measure meets criteria for public reporting.

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

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

ASCO's measure development process is rigorous, evidence-based, and utilizes the clinical expertise of multiple standing multi-disciplinary Technical Expert Panels (TEPs) dedicated to development and maintenance of measures across the cancer continuum. During measure maintenance, TEP members are provided with full measure specifications, applicable evidence, historical measure performance data, and any external feedback or requests for clarification or updates that have been received for the measure.

Staff on ASCO's measure development team are available to receive comments and questions from measure implementers and clinicians reporting the measures. As comments and questions are received, they are shared with appropriate staff for follow up. If comments or questions require expert input, these are shared with ASCO's TEPs to determine if measure modifications may be warranted. Additionally, for ASCO measures included in federal reporting programs, there is a system that has been established to elicit timely feedback and responses from ASCO staff in consultation with TEP members, as appropriate.

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.

See description in 4a2.1.1 above.

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.

In addition to the feedback obtained from a multi-disciplinary technical expert panel during the measure development and maintenance process, ASCO obtains feedback and receives measure inquiries from implementers and reporters via email. No specific feedback has been received by ASCO on this measure.

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

No specific feedback has been received by ASCO on this measure. However, we will continue to solicit feedback from MIPS users, and from the general public as we perform maintenance on this measure.

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

No additional feedback has been received by ASCO on this measure. However, we will continue to solicit feedback as we perform maintenance on this measure.

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

As stated in 4a2.2, ASCO did not receive specific feedback on this measure; therefore, ASCO's TEP did not consider external feedback from those being measured during revision of measure specifications or implementation.

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.

An analysis of MIPS data from 2017 at the TIN level indicated that while a slight majority (approximately 54%) of practices perform at 100% there are meaningful differences in performance across practices. Multiple practices perform at lower levels with the lowest performance score at 0% and average performance of 76% indicating room for improvement in a significant portion of practices. It should be noted that small sample size may impact the results presented, as the median denominator is 3, meaning that half of the performance in the graph above are based 3 patients or less.

Performance data from MIPS data 2017 does not include data for expanded RAS testing as those changes were implemented in 2018. We do not believe that the measure has been substantively changed in regard to its impact on reliability and validity as the data fields used and the clinical work flow remain the same; however, we do anticipate a greater performance gap due to the guideline update, which is a relatively new requirement in the field.

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.

At this time, we are not aware of any unintended consequences related to this measure. We take unintended consequences very seriously and therefore continuously monitor to identify actions that can be taken to mitigate them.

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

We have not observed any unexpected benefits associated with 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)

1860 : Patients with metastatic colorectal cancer and RAS gene mutation spared treatment with anti-epidermal growth factor receptor monoclonal antibodies

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

5a. Harmonization of Related Measures

The measure specifications are harmonized with related measures;

OR

The differences in specifications are justified

5a.1. If this measure conceptually addresses EITHER the same measure focus OR the same target population as NQF-endorsed measure(s):

Are the measure specifications harmonized to the extent possible?

Yes

5a.2. If the measure specifications are not completely harmonized, identify the differences, rationale, and impact on interpretability and data collection burden.

N/A - The measure specifications are harmonized.

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.)

An environmental scan did not identify competing measures. ASCO believes that NQF 1860 is a complementary measure assessing the inverse of the quality action captured in NQF 1859.

Appendix

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

No appendix **Attachment:**

Contact Information

Co.1 Measure Steward (Intellectual Property Owner): American Society of Clinical Oncology

Co.2 Point of Contact: Angela, Kennedy, Angela.Kennedy@asco.org, 571-483-1656-

Co.3 Measure Developer if different from Measure Steward: American Society of Clinical Oncology

Co.4 Point of Contact: Angela, Kennedy, Angela.Kennedy@asco.org

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.

ASCO's Gastrointestinal Technical Expert Panel (TEP) is a standing multi-disciplinary panel responsible for maintenance and de novo development of ASCO gastrointestinal measures. TEP members provide clinical expertise and guidance on measure concepts, level and quality of evidence, and measure specifications.

The current TEP roster is as follows:

David Ryan, MD (Chair), Massachusetts General Hospital

Nancy Baxter, Md, FRCSC, FACS, PhD, St. Michael's Hospital, University of Toronto

Emily Bergsland, MD, University of California, San Francisco

Jordan Berlin, MD, FASCO, Vanderbilt-Ingram Cancer Center

Philip Gold, MD, Swedish Cancer Institute

Theodore Hong, MD, Massachusetts General Hospital

Najjia Mahmoud, MD, University of Pennsylvania

Kim-Son Nguyen, MD, MPA, Palo Alto Medical Foundation / Sutter Health

Dan Zuckerman, MD, FASCO, St. Luke's Mountain States Tumor Institute

Measure Developer/Steward Updates and Ongoing Maintenance

Ad.2 Year the measure was first released: 2009

Ad.3 Month and Year of most recent revision: 07, 2019

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

Ad.5 When is the next scheduled review/update for this measure? 07, 2020

Ad.6 Copyright statement: The Measure is not clinical guidelines, does not establish a standard of medical care, and has not been tested for all potential applications.

The Measure, 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 Measure into a product or service that is sold, licensed or distributed for commercial gain.

Commercial uses of the Measure require a license agreement between the user and the American Society of Clinical Oncology (ASCO) and American Medical Association (AMA), [on behalf of the Physician Consortium for Performance Improvement® (PCPI®)] and prior written approval of ASCO, AMA, or PCPI. Neither ASCO, AMA, or PCPI, nor its members shall be responsible for any use of the Measures.

The AMA's and PCPI's significant past efforts and contributions to the development and updating of the Measure is acknowledged. ASCO is solely responsible for the review and enhancement ("Maintenance") of the Measures as of January 2015.

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

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

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

Ad.7 Disclaimers: See copyright statement in Ad.6 above.

Ad.8 Additional Information/Comments: