



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 #: 1858

Corresponding Measures:

De.2. Measure Title: Trastuzumab administered to patients with AJCC stage I (T1c) – III human epidermal growth factor receptor 2 (HER2) positive breast cancer who receive adjuvant chemotherapy

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

De.3. Brief Description of Measure: Percentage of female patients aged 18 and over with HER2/neu positive invasive breast cancer who are administered trastuzumab

1b.1. Developer Rationale: Approximately 15% of patients with breast cancer have tumors that overexpress the human epidermal growth hormone receptor protein (HER2). The American Society of Clinical Oncology (ASCO) envisions that use of this measure will improve concordance with recommendations for Trastuzumab administration for patients with AJCC stage I(T1c) – III, HER2/neu positive breast cancer. We recognize the importance of ensuring that the appropriate patient population receives guideline concordant treatment as studies have shown that the administration of Trastuzumab significantly improves overall survival in patients with high-risk HER2 positive breast cancer.

S.4. Numerator Statement: Patients for whom trastuzumab is administered within 12 months of diagnosis

S.6. Denominator Statement: Female patients aged 18 and over with AJCC stage I (T1c) – III, HER2/neu positive breast cancer who receive chemotherapy

S.8. Denominator Exclusions: Denominator Exclusions:

- o Patient transfer to practice after initiation of chemotherapy

Denominator Exceptions:

- o Reason for not administering trastuzumab documented (e.g. patient declined, patient died, patient transferred, contraindication or other clinical exclusion, neoadjuvant chemotherapy or radiation therapy not complete)

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:** Sep 16, 2015

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? | <input checked="" type="checkbox"/> Yes | <input type="checkbox"/> No |
| • Quality, Quantity and Consistency of evidence provided? | <input checked="" type="checkbox"/> Yes | <input type="checkbox"/> No |
| • Evidence graded? | <input checked="" type="checkbox"/> Yes | <input type="checkbox"/> No |

Summary of prior review in 2016

- The developer provided evidence that was based on two clinical practice guidelines:
 - American Society of Clinical Oncology (ASCO) guideline on systemic therapy for patients with advanced cancer: Clinicians should recommend HER2-targeted therapy–based combinations for first-line treatment, except for highly selected patients with ER-positive or PgR-positive and HER2-positive disease, for whom clinicians may use endocrine therapy alone. This guideline includes additional recommendations. **Strength of recommendation:** Strong. Evidence Quality: High
 - Cancer Care Ontario (CCO) guideline on optimal systemic therapy for early breast cancer in women: Trastuzumab plus chemotherapy is recommended for all patients with her2-positive, node-positive breast cancer and for patients with her2-positive, node-negative breast cancer greater than 1 cm in size. **CCO uses a narrative approach to grade the strength of recommendations. No additional details are provided regarding the grading.**
- The developer provided a systematic review of the evidence for the ASCO guideline and included Quantity, Quality, and Consistency of the evidence. No relevant studies had been conducted and published since the systematic reviews.
- The developer noted some contraindications to the HER2-targeted therapy due to its cardiovascular toxicity effects.

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 one additional clinical practice guideline:
 - The National Comprehensive Cancer Network (NCCN) guidelines on breast cancer: The panel recommends HER2-targeted therapy in patients with HER2-positive tumors. Trastuzumab is humanized monoclonal antibody with specificity for the extracellular domain of HER2. All of the adjuvant trials of trastuzumab have demonstrated clinically significant improvements in DFS, and the combined analysis from the NSABP B31 and NCCTG N9831 trials, and the HERA trial, showed significant improvement in OS with the use of trastuzumab in patients with high-risk, HER2-positive breast cancer. Therefore, regimens from each of these trials are included as trastuzumab-containing adjuvant regimen choices in the guideline. The benefits of trastuzumab are independent of ER status. Based on these studies, the panel has designated use of trastuzumab with chemotherapy as a category 1 recommendation in patients with HER2-positive tumors greater than 1 cm. **Evidence quality: High.**
- The developer provided a systematic review of the evidence for the ASCO guideline noting that a 2018 guideline update reaffirmed the recommendation of this measure. No new studies changed the conclusions reached by the 2018 guideline.
- The developer provided a systematic review of the evidence for the CCO guideline noting that updated guidelines continue to support the measure.

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

Process measure/systematic review (Box 3) → QQC provided for NCCN guideline (Box 4) → Systematic review concluded: Quantity: High; Quality: High; Consistency: High (Box 5a) → High

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

RATIONALE:

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 data provided by CMS:
Number of unique entities: Frequency 73

Denominators

Min: 1; Q1: 2; Median: 5; Q3: 23; Max: 206; Total: 1815

Measure Distribution

Min: 0; Q1: 0.9853; Median: 1; Mean: 0.9307; Q3: 1; Max: 1; CI for mean: (0.89, 0.98); Percent outside CI: 90.41

An analysis of 250 unique NPIs indicated results similar to the TIN-level analysis, in that many NPIs have a small denominator, and the majority are already performing at 100 percent. Additional details from the NPI-level analysis are provided below.

Unique Number of NPIs: 250

Distribution of Measure Denominators and Measure Performance:

Denominator

Min: 0; Q1: 2; Median: 3; Mean: 6.072; Q3: 7; Max: 45

Measure:

Min: 0; Q1: 1; Median: 1; Mean: 0.9206; Q3: 1; Max: 1; CI for mean: (0.89, 0.95); Percent outside CI: 97.19

- 2017 QPP Experience Report Appendix indicates performance on this measure is 97.51%. 2019 benchmarking data for QI 450 indicates a topped out measure.

Disparities

No disparities data was presented. However, the developer cited a 2018 systematic review and meta-analysis that notes that the uptake of trastuzumab therapy is widely variable between studies and across subgroups, suggesting disparities in the use of trastuzumab.

Questions for the Committee:

- Does the Committee agree that the measure is topped out?
- 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: This measure is topped out.

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.”

- Good evidence
- process measure - but well defined path to quality of care
- Evidence is strong supporting use of trastuzumab in HER2 positive breast CA

- The evidence provided relates directly to the process/outcome being measured with regards to patients being treated with Trastuzumab and in accordance to ASCO guidelines.
- Patients would value this
- Strong evidence for measure from several trials.
- This is a process measure based on clinical guidelines recommendation. NCCN guideline recommendations are based on high level evidence (Level 1). The developers provided guidelines and multiple systemic reviews. The guidelines were updated in 2018. There is ongoing evidence from large RCT that the use of trastuzumab in HER2+ disease is associated with improved DFS and OS. The evidence for this measure is strong.
- Evidence was updated from last submission and relates directly to the metric
- Strong high quality evidence is presented.
- I agree with the preliminary rating of "High" and the committee should probably vote on the new evidence.

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?

- This may be topped out
- Gap is fairly small, but it exists. would agree with topped out
- Data were provided, showing near 100% compliance c/w topped out measure
- Yes, data presented and 2017 MIPS performance data provided by CMS.
- Very little performance gap.
- The data provided suggests near perfect performance by providers and it is hard to see any gaps. Additionally treatment of breast cancer is moving toward more neoadjuvant therapy and relevance of this measure is decreasing.
- The developer used MIPS data from 2017 which indicated a performance on this measure is 97.51%. 2019 benchmarking data indicates a topped out measure. No disparities data was available with the MIPS data. However, the developers included information from a 2018 systematic review and meta-analysis of observational studies by Martin, et al., which identified large variability in uptake of trastuzumab in HER2-positive early breast cancer patients (9.1-100%) and metastatic breast cancer patients (50.8-84.0%), with a pooled uptake of 71.3%. The developers clarified that this data is heterogeneous and should be interpreted with caution. Since the literature suggests an ongoing wide variation in administration of trastuzumab to appropriate patients, this measure has moderate importance to measure.
- Performance gap is low and appears almost topped out.
- The performance gap is very low with many already performing at 100%
- I agree with the preliminary rating of "Low".

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

Questions for the Committee regarding reliability:

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

Questions for the Committee regarding validity:

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

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

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

Scientific Acceptability: Preliminary Analysis Form

Measure Number: 1858

Measure Title: Trastuzumab administered to patients with AJCC stage I (T1c) – III human epidermal growth factor receptor 2 (HER2) positive breast cancer who receive adjuvant chemotherapy

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-noise 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 0.9657 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.

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

Submission document: Testing attachment, section 2b4.

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 conducted a Pearson correlation analysis to determine the association between performance scores of the shared providers.
- The developer interpreted 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 0.711, 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.

Exclusion analysis not provided; however, rationale was provided. Developer identified statistically significant and meaningful differences (Box 1) → Empirical validity testing was conducted (Box 2) → Validity testing was conducted for provider level entity (Box 5) → Correlation of performance measure scores conducted and reported (Box 6) → Strong, positive correlation reported (7a) → 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?

- Good
- no issues- site v. practice is unlikely to change answers

- provided testing at wrong level, but overall components reliable.
- I do not have any concerns about whether the measure can be consistently implemented.
- no concerns of reliability
- No reliability issues
- The data elements are clearly defined. The measure description is complete and concise. I believe that this measure can be consistently implemented.
- Data elements are defined clearly. The only concern would be the pitfalls of manually abstracted measurement.
- High reliability with mean of .9657
- No concerns but probably worth committee discussion.

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

- No
- no
- No
- I do not have any concerns.
- No
- No concern
- 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. The developers reported a mean reliability of 0.9657 which is considered very high according to Adams' definition. If the developers can clarify the level of testing, this represents a high level of reliability. The NQF staff rated it as insufficient.
- Yes
- No concerns, high reliability documented.
- I would like the committee to discuss the preliminary rating of "Insufficient".

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

- No
- no
- No
- I do not have any concerns.
- None
- No issues
- I have no concerns. Testing was done on the performance measure score. The developers correlated two measures that were positively associated—AJCC stage 1 (T1c) – III HER2 positive breast cancer patients who receive chemotherapy and are administered trastuzumab (NQF #1858/QI 450) and HER2 negative or undocumented breast cancer patients spared treatment with HER2-targeted therapies (NQF #1859/QI 449). This is a reasonable hypothesis. The correlation was 0.711 indicating a strong, positive correlation between performance scores of the shared providers. Face validity testing also demonstrated a vast majority of respondents (95%) strongly agree or agree that the measure provided an accurate reflection of quality and can be used to distinguish good and poor quality.

- No
- No concerns high validity demonstrated
- No concerns

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?

- Denominators too small for many
- failure to do this without cause would be a significant variation in care
- No threats, but if measure is topped out it is unlikely to drive quality.
- I believe analysis provides comparable results, especially in conjunction with ASCO
- NO concerns
- The measure is drawn from registry and medical records. There is a slim possibility that the HER2 status are not captured/mis captured on the records. This small erroneous capturing of HER2 should not affect this measure. likelihood of missing date is slim.
- There are no threats to the validity. There were no risk adjustments and the rate of exclusions was not presented. The testing appears to support the ability to detect meaningful differences however the measure does appear to be topped out. 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. The majority of TINs perform at 100%, although there are multiple TINs whose performance is below 25%. There was no missing data in the MIPS data.
- N/A
- No
- 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 risk adjustment
- no issues
- exclusions are appropriate.
- I do not feel any patients or patient groups are inappropriately excluded from the measure and believe that the appropriate risk adjustment strategy is included.

- No threats I see.
- N/A
- There are no risk adjustments. Exclusion rates are not presented.
- N/A
- 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?

- It's feasible
- should be routine
- no feasibility concerns
- I don't have any concerns at this time.
- None

- The issue of adjuvant therapy is the central part of this measure. Adjuvant therapy is treatment after surgery in curable patients. Inaccurate capturing of neoadjuvant therapy may cause problem in case ascertainment.
- The data elements are routinely generated and used during the delivery of care. Most of the data elements are not in electronic form. To acquire the data, chart audits of the medical record or cancer registry data is required. The measure has been in use for many years and has been demonstrated to be feasible. I have no concerns about the ability to put this measure into operational use.
- 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.
- No concerns about feasibility
- 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

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 of 97.51% in the 2017 QPP Experience Report Appendix. 2019 MIPS benchmarking data for QI 450 indicates this measure as topped out.

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 unintended 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?

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?

- It's usable
- mips measure
- in use in accountability program. no specific feedback.
- The measure provides greater benefit then harm.
- Yes
- Reporting is through MIPS and physician compare reports. I am not sure about discussing the MIPS results of the practice and individual with providers
- The measure is in use for QOPI and MIPS. The measure reflects the standard of care for HER2+ patients and has been used as intended.
- Measure is used in multiple reporting programs
- No concerns, publicly reported in MIPS and ASCO
- 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.

- No harms
- none
- no unintended consequences, but topped out.
- The performance results can improve the goal of high quality.
- Good benefit vs. harms
- This measure seems to be topped out.
- The measure reflects the standard of care for HER2+ patients and the use of trastuzumab in these patients has been associated with improved DFS and OS. The benefits of this measure far outweigh any potential risks. The measure can be used to improve performance. The measure appears to be topped out but there are likely opportunities for performance improvement given the range of performance noted in the literature.
- Metric is topped out and not much room for improvement, no evidence of unintended consequences
- The measure is already performing very well with minimal room for additional improvement in high quality healthcare.
- I agree with the "Moderate" preliminary rating.

Criterion 5: [Related and Competing Measures](#)

Related or competing measures

1855 Quantitative HER2 evaluation by IHC uses the system recommended by the ASCO/CAP guidelines

1857 HER2 negative or undocumented breast cancer patients spared treatment with HER2-targeted therapies

Harmonization

No harmonization issues; 1855 and 1857 are complementary measures to 1858

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?

- Related measures are compatible
- complementary
- no
- No additional steps necessary at this time.
- Not that I know of
- 1855 and 1857 are related but not competing.
- There are related measures that have been harmonized.
- No
- There are complimentary measures but no competing measures.
- I agree with the assessment that 1855 & 1857 are complementary.

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

[1858_Evidence_MSF5.0_Data_11.23.2019.doc,NQF_evidence_attachment_11.23.2019-637102982946754578.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): 1858

Measure Title: [Trastuzumab administered to patients with AJCC stage I \(T1c\) – III human epidermal growth factor receptor 2 \(HER2\) positive breast cancer who receive adjuvant chemotherapy](#)

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

Date of Submission: [11/12/2019](#)

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](#)

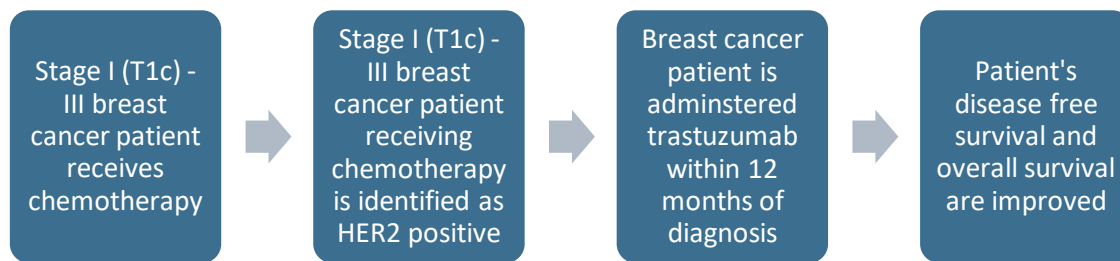
☒ Process: [Administration of trastuzumab](#)

☐ Appropriate use measure: [Click here to name what is being measured](#)

☐ Structure: [Click here to name the structure](#)

☐ Composite: [Click here to name what is being measured](#)

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



The process evaluated in this measure is a patient being administered trastuzumab within 12 months of a breast cancer diagnosis. Multiple randomized controlled trials have demonstrated that administration of trastuzumab improves a patient's disease-free survival (DFS) and overall survival (OS). Additionally, this measure is directly supported by recommendations in National Comprehensive Cancer Network (NCCN), Cancer Care Ontario (CCO), and American Society of Clinical Oncology(ASCO)-CCO clinical practice guidelines.

The role of trastuzumab in adjuvant and neoadjuvant therapy in women with HER2/neu-overexpressing breast cancer. Madarnas Y, Tey R, reviewers. Toronto (ON): Cancer Care Ontario; 2011 Sep 15 [Endorsed 2010 Jun 11]. Program in Evidence-based Care Evidence-Based Series No.: 1-24 Version 2
<https://www.cancercare.on.ca/common/pages/UserFile.aspx?fileId=13890>

Gradishar WJ, Anderson BO, Abraham J, et al. NCCN Guidelines Panel. NCCN Clinical Practice Guidelines in Oncology – Breast Cancer. Version 3. 2019. September 6, 2019.
<https://www.nccn.org> (free account is required to view guideline)

Denduluri, N., et al., *Selection of Optimal Adjuvant Chemotherapy and Targeted Therapy for Early Breast Cancer: ASCO Clinical Practice Guideline Focused Update*. J Clin Oncol, 2018. **36**(23): p. 2433-2443.
<https://www.asco.org/practice-guidelines/quality-guidelines/guidelines/breast-cancer#/11081>

Eisen A, Fletcher GG, Gandhi S, Mates M, Freedman OC, Dent SF, et al. Optimal systematic therapy for early female breast cancer. Toronto (ON): Cancer Care Ontario; 2014 Sep 30 [In Review 2019 Jan]. Program in Evidence-Based Care Evidence-Based Series No.: 1–21 IN REVIEW.
<https://www.cancercareontario.ca/en/guidelines-advice/types-of-cancer/331>

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: <ul style="list-style-type: none">TitleAuthorDateCitation, including page numberURL	NCCN Guidelines Version 3.2019 Breast Cancer National Comprehensive Cancer Network Version 3.2019 – September 6, 2019 NCCN Clinical Practice Guidelines in Oncology™. Breast Cancer, V.3.2019 (MS-30) https://www.nccn.org (free account is required to view the guideline, however full pdf is attached below)  NCCN breast guideline.pdf
Quote the guideline or recommendation verbatim about the	“The panel recommends HER2-targeted therapy in patients with HER2-positive tumors. Trastuzumab is humanized monoclonal antibody with specificity for the extracellular domain of HER2. All of the adjuvant trials of trastuzumab have demonstrated clinically significant improvements in DFS, and the combined

process, structure or intermediate outcome being measured. If not a guideline, summarize the conclusions from the SR.	analysis from the NSABP B31 and NCCTG N9831 trials, and the HERA trial, showed significant improvement in OS with the use of trastuzumab in patients with high-risk, HER2-positive breast cancer. Therefore, regimens from each of these trials are included as trastuzumab-containing adjuvant regimen choices in the guideline. The benefits of trastuzumab are independent of ER status. Based on these studies, the panel has designated use of trastuzumab with chemotherapy as a category 1 recommendation in patients with HER2-positive tumors greater than 1 cm.” (MS-44-MS-46)
Grade assigned to the evidence associated with the recommendation with the definition of the grade	Category 1: Based upon high-level evidence, there is uniform NCCN consensus that the intervention is appropriate
Provide all other grades and definitions from the evidence grading system	<p>NCCN Categories of Evidence and Consensus:</p> <p>Category 1: Based upon high-level evidence, there is uniform NCCN consensus that the intervention is appropriate.</p> <p>Category 2A: Based upon lower-level evidence, there is uniform NCCN consensus that the intervention is appropriate.</p> <p>Category 2B: Based upon lower-level evidence, there is NCCN consensus that the intervention is appropriate.</p> <p>Category 3: Based upon any level of evidence, there is major NCCN disagreement that the intervention is appropriate.</p>
Grade assigned to the recommendation with definition of the grade	Category 1: Based upon high-level evidence, there is uniform NCCN consensus that the intervention is appropriate
Provide all other grades and definitions from the recommendation grading system	<p>NCCN Categories of Evidence and Consensus:</p> <p>Category 1: Based upon high-level evidence, there is uniform NCCN consensus that the intervention is appropriate.</p> <p>Category 2A: Based upon lower-level evidence, there is uniform NCCN consensus that the intervention is appropriate.</p>

	<p>Category 2B: Based upon lower-level evidence, there is NCCN consensus that the intervention is appropriate.</p> <p>Category 3: Based upon any level of evidence, there is major NCCN disagreement that the intervention is appropriate.</p>
<p>Body of evidence:</p> <ul style="list-style-type: none"> Quantity – how many studies? Quality – what type of studies? 	<p>The NCCN guideline notes that results of nine randomized trials testing trastuzumab as adjuvant therapy have been reported, and recounts detailed results for four RCTs of adjuvant trastuzumab, including NSABP B-31, NCCTG N9831, HERA, and BCIRG 006. NCCN’s analysis of these trials includes the following summary (MS-44-MS-46):</p> <ul style="list-style-type: none"> “The panel recommends HER2-targeted therapy in patients with HER2-positive tumors. Trastuzumab is humanized monoclonal antibody with specificity for the extracellular domain of HER2. All of the adjuvant trials of trastuzumab have demonstrated clinically significant improvements in DFS, and the combined analysis from the NSABP B31 and NCCTG N9831 trials, and the HERA trial, showed significant improvement in OS with the use of trastuzumab in patients with high-risk, HER2-positive breast cancer. Therefore, regimens from each of these trials are included as trastuzumab-containing adjuvant regimen choices in the guideline. The benefits of trastuzumab are independent of ER status. Based on these studies, the panel has designated use of trastuzumab with chemotherapy as a category 1 recommendation in patients with HER2-positive tumors greater than 1 cm.”
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.

<p>Source of Systematic Review:</p> <ul style="list-style-type: none"> Title Author 	<p>Optimal Systemic Therapy for Early Female Breast Cancer</p> <p>Andrea Eisen, Glenn G. Fletcher, Sonal Gandhi, Mihaela Mates, Orit C. Freedman, Susan F. Dent, Maureen E. Trudeau, and members of the Early Breast Cancer Systemic Therapy Consensus Panel</p> <p>September 30, 2014</p>
--	---

<ul style="list-style-type: none"> • Date • Citation, including page number • URL 	<p>Eisen A, Fletcher GG, Gandhi S, Mates M, Freedman OC, Dent SF, et al. Optimal systematic therapy for early female breast cancer. Toronto (ON): Cancer Care Ontario; 2014 Sep 30 [In Review 2019 Jan]. Program in Evidence-Based Care Evidence-Based Series No.: 1–21 IN REVIEW. (pgs. 17-18)</p> <p>https://www.cancercareontario.ca/en/guidelines-advice/types-of-cancer/331</p>
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.	R27. Trastuzumab plus chemotherapy is recommended for all patients with HER2+ node positive breast cancer and for patients with Her2+ node negative breast cancer greater than 1 cm in size
Grade assigned to the evidence associated with the recommendation with the definition of the grade	<p>This guideline utilized a modified Delphi technique to reach consensus on final recommendations.</p> <p>The Program in Evidence-Based Care (PEBC) is an initiative of the Ontario provincial cancer system, Cancer Care Ontario (CCO). The PEBC produces evidence-based and evidence-informed guidelines, known as Evidence-Based Series (EBS) reports, using the methods of the Practice Guidelines Development Cycle. The EBS report consists of an evidentiary base (typically a systematic review), an interpretation of and consensus agreement on that evidence by our Groups or Panels, the resulting recommendations, and an external review by Ontario clinicians and other stakeholders in the province for whom the topic is relevant.</p>
Provide all other grades and definitions from the evidence grading system	See Body of Evidence section.
Grade assigned to the recommendation with definition of the grade	See Body of Evidence section.
Provide all other grades and definitions from the	See Body of Evidence section.

recommendation grading system	
<p>Body of evidence:</p> <ul style="list-style-type: none"> Quantity – how many studies? Quality – what type of studies? 	<p>Key Evidence and Qualifying Statements</p> <ul style="list-style-type: none"> Phase III clinical studies have demonstrated improved DFS and OS with the addition of trastuzumab to chemotherapy compared with chemotherapy alone in HER2+ early breast cancer (see Table 14 for Evidentiary Base). The majority of adjuvant trastuzumab trials included patients with lymph node positive breast cancer, or lymph node negative disease with one of the following high-risk features: ER-, grade 2 or 3, T \geq1cm, or age <35 years. Trastuzumab may still be considered in patients with HER2+ disease outside these features. Although most studies excluded patients with tumors <1 cm, the benefit of trastuzumab was equivalent in both node negative and node positive tumors in the HERA trial which included small N0 tumours (1 cm was the formal inclusion criteria, although 60 patients with tumors <1 cm were also enrolled). The BCIRG 006 trial analysis by tumour size found benefit in tumours <1 cm, <2 cm, and \geq2 cm, but not for tumours 1-2 cm in size; however, interpretation is limited because of the small number of patients in each category. The review by Petrelli and Barni concluded that patients with HER2+ tumours have a higher rate of recurrence and poorer survival rate than patients with HER2- cancer of the same size/stage, confirming that HER2 positivity itself is a risk factor. There does not appear to be a threshold according to tumour size, and size alone should not be the deciding factor in whether to administer trastuzumab to patients with tumours <1 cm. In Ontario, tumours <1 cm can be treated under the Evidence Building Program (EBP). The meta-analysis by Moja et al (Cochrane Collaboration) found that the hazard ratio for trastuzumab-containing regimens vs. chemotherapy alone was 0.66 for OS and 0.60 for DFS ($p < 0.00001$ for both). The risk of congestive heart failure and left ventricular ejection decline were higher with trastuzumab (RR=55.1, $p < 0.00001$ and R=1.83, $p < 0.0008$, respectively). In patients at high risk of recurrence without cardiac problems, there is clear survival rate benefit for trastuzumab.
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	Updated guidelines continue to support this measure.

since the SR. Do the new studies change the conclusions from the SR?	
--	--

<p>Source of Systematic Review:</p> <ul style="list-style-type: none"> • Title • Author • Date • Citation, including page number • URL 	<p>The role of trastuzumab in adjuvant and neoadjuvant therapy in women with HER2/neu-overexpressing breast cancer. Madarnas Y, Tey R, reviewers. Toronto (ON): Cancer Care Ontario; 2011 Sep 15 [Endorsed 2010 Jun 11]. Program in Evidence-based Care Evidence-Based Series No.: 1-24 Version 2</p> <p>https://www.cancercareontario.ca/sites/ccocancercare/files/guidelines/full/pebc1-24f.pdf (a pop-up box will appear, click "OK")</p> <div data-bbox="435 793 483 856" data-label="Image"> </div> <p>CCO breast guideline.pdf</p> <p>Please note the verbatim recommendation (below) appeared in the above Cancer Care Ontario guideline, which is no longer available via PubMed. This recommendation was reaffirmed in the following 2018 ASCO and Cancer Care Ontario guideline update:</p> <p>Selection of Optimal Adjuvant Chemotherapy and Targeted Therapy for Early Breast Cancer</p> <p>Denduluri, N., et al., <i>Selection of Optimal Adjuvant Chemotherapy and Targeted Therapy for Early Breast Cancer: ASCO Clinical Practice Guideline Focused Update</i>. J Clin Oncol, 2018. 36(23): p. 2433-2443.</p> <p>https://www.asco.org/practice-guidelines/quality-guidelines/guidelines/breast-cancer#/11081</p>
Quote the guideline or recommendation verbatim about the process, structure or intermediate outcome being measured. If not a guideline,	<p>"Trastuzumab should be offered for one year to all patients with HER2 Positive node-positive or node-negative, tumour greater than 1 cm in size, and primary breast cancer and who are receiving or have received (neo)adjuvant chemotherapy. Trastuzumab should be offered after chemotherapy." (CCO guideline, development and methods pg 3/ pdf pg 29;</p> <p>https://www.cancercareontario.ca/sites/ccocancercare/files/guidelines/full/pebc1-24f.pdf (a pop-up box will appear, click "OK").</p>

summarize the conclusions from the SR.	Please note this original recommendation was reaffirmed in a 2018 ASCO and Cancer Care Ontario guideline update, available at: https://www.asco.org/practice-guidelines/quality-guidelines/guidelines/breast-cancer#/11081
Grade assigned to the evidence associated with the recommendation with the definition of the grade	CCO guidelines use a narrative approach in grading the quality of the evidence.
Provide all other grades and definitions from the evidence grading system	CCO guidelines use a narrative approach in grading the quality of the evidence.
Grade assigned to the recommendation with definition of the grade	<p>The guideline provides strong support for the use of trastuzumab in all patients with HER2 positive primary breast cancer.</p> <p>Strong Recommendation: There is high confidence that the recommendation reflects best practice. This is based on (1) strong evidence for a true net effect (e.g., benefits exceed harms); (2) consistent results, with no or minor exceptions; (3) minor or no concerns about study quality; and/or (4) the extent of panelists' agreement. Other compelling considerations (discussed in the guideline's literature review and analyses) may also warrant a strong recommendation.</p>
Provide all other grades and definitions from the recommendation grading system	<p>Strong Recommendation: There is high confidence that the recommendation reflects best practice. This is based on (1) strong evidence for a true net effect (e.g., benefits exceed harms); (2) consistent results, with no or minor exceptions; (3) minor or no concerns about study quality; and/or (4) the extent of panelists' agreement. Other compelling considerations (discussed in the guideline's literature review and analyses) may also warrant a strong recommendation.</p> <p>Moderate Recommendation: There is moderate confidence that the recommendation reflects best practice. This is based on (1) good evidence for a true net effect (e.g., benefits exceed harms); (2) consistent results, with minor and/or few exceptions; (3) minor and/or few concerns about study quality; and/or (4) the extent of panelists' agreement. Other compelling considerations (discussed in the guideline's literature review and analyses) may also warrant a moderate recommendation.</p>

	<p>Weak Recommendation: There is some confidence that the recommendation offers the best current guidance for practice. This is based on (1) limited evidence for a true net effect (e.g., benefits exceed harms); (2) consistent results, but with important exceptions; (3) concerns about study quality; and/or (4) the extent of panelists' agreement. Other considerations (discussed in the guideline's literature review and analyses) may also warrant a weak recommendation.</p>
<p>Body of evidence:</p> <ul style="list-style-type: none"> • Quantity – how many studies? • Quality – what type of studies? 	<p>Six randomized controlled trials were considered in the original CCO recommendation.</p>
<p>Estimates of benefit and consistency across studies</p>	<p>The evidence described in the RCTs is directly relevant to the measure (use of trastuzumab in patients with HER2/neu positive breast cancer). All studies considered women with invasive breast cancer that overexpressed HER2/neu and outcomes associated with the inclusion of trastuzumab. Five trials included chemotherapy plus or minus trastuzumab. Some of those also investigated the schedule for trastuzumab delivery; considering schedules concurrent with chemotherapy, or following chemotherapy completion with various time periods in between completion of chemotherapy and the start of trastuzumab. Duration of trastuzumab was also investigated. One trial specifically considered cardiac adverse events to assess potential harms, other trials considered adverse events in addition to disease-specific outcomes.</p> <p>The outcome of disease-free survival is more precise given the limited long-term follow-up, with more events for consideration, compared to overall survival. This limits issues with insufficient events. Notably, both outcomes were reported for consideration, though benefits in overall survival were noted in two individual studies and the combined analysis from NSABP B31 and NCCTG N9831.</p> <p>Results were consistent with respect to improvements in disease-free survival among women randomized to trastuzumab-containing arms. Hazard ratios reported for disease free survival were 0.54, 0.55, 0.45 and 0.48. The results for disease-free survival across trials were statistically significant for the treatment arm including trastuzumab.</p>

	<p>Studies were consistent with respect to the direction of effect. Differences in magnitude were noted, but can be attributed to various chemotherapies regimens across the studies, as well as slightly different patient populations.</p> <p>All of the adjuvant trials of trastuzumab have demonstrated clinically significant improvements in disease-free survival. The combined analysis from NSABP B31 and NCCTG N9831, BCIRG 006, and the HERA trial showed significant improvement in overall survival with the use of trastuzumab in patients with high-risk, HER2 positive breast cancer.</p> <p>Based on preliminary reports of three large RCTs, the addition of one year of trastuzumab, following a variety of adjuvant or neoadjuvant chemotherapy regimens, significantly improved the primary endpoint of DFS in patients with HER2/neu positive early breast cancer. Secondary endpoints of RFS, DDFS, and TTR in all studies, and OS in one combined study, were also significantly improved with the addition of trastuzumab. Those results are only applicable to women with HER2/neu overexpressing breast cancer who complete a minimum of four cycles of adjuvant or neoadjuvant chemotherapy. Although the majority of the patients in those studies had node-positive breast cancer, women with high-risk node-negative breast cancer were also included in HERA (32% were N0 but had T1c tumors) and NCCTG 9831 (11% were N0 but had tumours >1cm if ER negative, >2cm if ER positive). Therefore, those results are also generalizable to women with node-negative breast cancer who meet those criteria. The magnitude of incremental benefit conveyed by adjuvant trastuzumab well exceeds the gains accrued by over three decades of adjuvant chemotherapy use.</p>
What harms were identified?	<p>Based on the current reports, the cardiac toxicity with adjuvant trastuzumab appears to be acceptable. Notably, the reported rate of cardiac events was higher in the concurrent versus sequential trastuzumab arm (in NSABP B31 4.1% vs. 0.7%, HR of 7.2; in NCCTG 9831 3.3% vs. 2.2%). The toxicity is considered acceptable, given the increase in survival.</p>
Identify any new studies conducted since the SR. Do the new studies	<p>A 2018 ASCO and Cancer Care Ontario guideline update reaffirmed this recommendation on the use of trastuzumab, following a systematic review. No new studies changing the conclusions reached by the 2018 guideline update were found in subsequent literature reviews.</p>

change the conclusions from the SR?	
-------------------------------------	--

1a.4 OTHER SOURCE OF EVIDENCE

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

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

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

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

1b. Performance Gap

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

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

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

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

Approximately 15% of patients with breast cancer have tumors that overexpress the human epidermal growth hormone receptor protein (HER2). The American Society of Clinical Oncology (ASCO) envisions that use of this measure will improve concordance with recommendations for Trastuzumab administration for patients with AJCC stage I(T1c) – III, HER2/neu positive breast cancer. We recognize the importance of ensuring that the appropriate patient population receives guideline concordant treatment as studies have shown that the administration of Trastuzumab significantly improves overall survival in patients with high-risk HER2 positive breast cancer.

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

2019 Submission

Testing to identify statistically significant and meaningful differences in performance was conducted using 2017 MIPS performance from registry data provided from CMS. Practices were identified by unique number of TINs, and the 2017 data was from 73 unique TINs. 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 during the 2017 program year. We were unable to determine from our rolled-up data sample the number of clinicians who reported to MIPS as an individual or a 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. The NPI-level analysis of the 2017 MIPS data was conducted on 254 denominator-eligible patients. Additional descriptive characteristics of the measured patients are unknown. Eligible patients were included in this data set according to the reporting requirements for the 2017 MIPS program year.

An analysis of 73 unique TINs indicated that 25 percent of TINs have a denominator of two or less, and 50 percent of TINs have a denominator of five or less, and that the measure is heavily skewed with a large proportion of TINs performing perfectly (roughly 55 out of the 73 TINs). Additional details from the TIN-level analysis are provided below.

Number of unique entities: Frequency 73

Denominators

Min: 1; Q1: 2; Median: 5; Q3: 23; Max: 206; Total: 1815

Measure Distribution

Min: 0; Q1: 0.9853; Median: 1; Mean: 0.9307; Q3: 1; Max: 1; CI.for.mean: (0.89, 0.98); Percent.outside.CI: 90.41

An analysis of 250 unique NPIs indicated results similar to the TIN-level analysis, in that many NPIs have a small denominator, and the majority are already performing at 100 percent. Additional details from the NPI-level analysis are provided below.

Unique Number of NPIs: 250

Distribution of Measure Denominators and Measure Performance:

Denominator

Min: 0; Q1: 2; Median: 3; Mean: 6.072; Q3: 7; Max: 45

Measure:

Min: 0; Q1: 1; Median: 1; Mean: 0.9206; Q3: 1; Max: 1; CI.for.mean: (0.89, 0.95); Percent.outside.CI: 97.19

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.

In a 2018 systematic review and meta-analysis of observational studies, Martin, et al. identified large variability in uptake of trastuzumab in HER2-positive early breast cancer patients (9.1-100%) and metastatic breast cancer patients (50.8-84.0%), with a pooled uptake of 71.3%. The authors noted the uptake of trastuzumab therapy varied widely between studies and across subgroups suggesting that there may be some inequalities in the use of trastuzumab.

Martin, A.P., et al., Trastuzumab uptake in HER2-positive breast cancer patients: a systematic review and meta-analysis of observational studies. Crit Rev Oncol Hematol, 2018. 130: p. 92-107.

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 and 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

In a 2018 systematic review and meta-analysis of observational studies, Martin, et al. identified large variability in uptake of trastuzumab in HER2-positive early breast cancer patients (9.1-100%) and metastatic breast cancer patients (50.8-84.0%), with a pooled uptake of 71.3%. The authors noted the uptake of trastuzumab therapy varied widely between studies and across subgroups, suggesting inequalities exist in the use of trastuzumab. The authors suggested a cautious interpretation of findings due to study heterogeneity and potential confounding, and recommended additional studies using individual level data controlled for confounders in order to gain a better understanding about inequalities in trastuzumab use.

Martin, A.P., et al., Trastuzumab uptake in HER2-positive breast cancer patients: a systematic review and meta-analysis of observational studies. Crit Rev Oncol Hematol, 2018. 130: p. 92-107.

2. Reliability and Validity—Scientific Acceptability of Measure Properties

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

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

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

Cancer, Cancer : Breast

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

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_450_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.

As 2017 MIPS data were used to complete updated testing, exclusions from the previous submission have been aligned with the MIPS specifications. We have also delineated between exceptions and exclusions in accordance with the MIPS measure specification.

Please note that one exclusion of “patient has metastatic disease at diagnosis” was removed from due to redundancy, as the denominator population is already limited to patients with stage I (T1c) - III cancer. We also intend to remove this from the MIPS specification as the measure update cycle allows.

We do not consider these modifications to be substantive to the measure.

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

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

Patients for whom trastuzumab is administered within 12 months of diagnosis

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

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

Numerator:

Trastuzumab administered within 12 months of diagnosis

Numerator Options:

Performance Met: Trastuzumab administered within 12 months of diagnosis

OR

Denominator Exception: Reason for not administering Trastuzumab documented (e. g. patient declined, patient died, patient transferred, contraindication or other clinical exclusion, neoadjuvant chemotherapy or radiation NOT complete)

OR

Performance Not Met: Trastuzumab not administered within 12 months of diagnosis

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

Female patients aged 18 and over with AJCC stage I (T1c) – III, HER2/neu positive breast cancer who receive chemotherapy

S.7. Denominator Details (All information required to identify and calculate the target population/denominator such as definitions, time period for data collection, specific data collection items/responses, code/value sets – Note: lists of

individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b.)

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

Denominator Criteria (Eligible Cases):

Female Patients aged = 18 years on date of encounter

AND

Diagnosis of breast cancer

AND

Patient encounter during performance period

AND

Two or more encounters at the reporting site AND

Breast Adjuvant Chemotherapy administered:

AND

HER-2/neu positive:

AND

AJCC stage at breast cancer diagnosis = II or III: G9831

OR

AJCC stage at breast cancer diagnosis = I (IA or IB) and T-Stage at breast cancer diagnosis does NOT equal = T1, T1a, T1b

AND NOT

Denominator Exclusions:

Patient transfer to practice after initiation of chemotherapy

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

Denominator Exclusions:

- o Patient transfer to practice after initiation of chemotherapy

Denominator Exceptions:

- o Reason for not administering trastuzumab documented (e.g. patient declined, patient died, patient transferred, contraindication or other clinical exclusion, neoadjuvant chemotherapy or radiation therapy not complete)

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.)*

Denominator Exclusions:

Patient transfer to practice after initiation of chemotherapy

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, no risk stratification

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 with exclusions and exceptions; thus, the calculation algorithm is: Patients meeting the numerator + patients with valid exceptions/ (Patients in the denominator – Patients with valid exclusions) 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, measure is not based on a survey or instrument

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.)

N/A

2. Validity – See attached Measure Testing Submission Form

1858_nqf_testing_attachment_7.31.2019.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): 1858

Measure Title: Trastuzumab administered to patients with AJCC stage I (T1c) – III human epidermal growth factor receptor 2 (HER2) positive breast cancer who receive adjuvant chemotherapy

Date of Submission: TBD

Type of Measure:

<input type="checkbox"/> Outcome (including PRO-PM)	<input type="checkbox"/> Composite – STOP – use composite testing form
<input type="checkbox"/> Intermediate Clinical Outcome	<input type="checkbox"/> Cost/resource
<input checked="" type="checkbox"/> Process (including Appropriate Use)	<input type="checkbox"/> Efficiency
<input type="checkbox"/> Structure	

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:
<input checked="" type="checkbox"/> abstracted from paper record	<input checked="" type="checkbox"/> abstracted from paper record
<input type="checkbox"/> claims	<input type="checkbox"/> claims
<input checked="" type="checkbox"/> registry	<input checked="" type="checkbox"/> registry
<input type="checkbox"/> abstracted from electronic health record	<input type="checkbox"/> abstracted from electronic health record
<input type="checkbox"/> eMeasure (HQMF) implemented in EHRs	<input type="checkbox"/> eMeasure (HQMF) implemented in EHRs
<input type="checkbox"/> other: Click here to describe	<input type="checkbox"/> 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 are from the fall 2011 QOPI round (reflecting data submitted October and November 2011) as well as 2017 MIPS performance data. The MIPS performance year begins on January 1 and ends December 31 of 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 all the levels specified and intended for measure implementation, e.g., individual clinician, hospital, health plan)

Measure Specified to Measure Performance of: (must be consistent with levels entered in item S.20)	Measure Tested at Level of:
<input type="checkbox"/> individual clinician	<input type="checkbox"/> individual clinician
<input checked="" type="checkbox"/> group/practice	<input checked="" type="checkbox"/> group/practice
<input type="checkbox"/> hospital/facility/agency	<input type="checkbox"/> hospital/facility/agency
<input type="checkbox"/> health plan	<input type="checkbox"/> health plan
<input type="checkbox"/> other: Click here to describe	<input type="checkbox"/> other: Click here to describe

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

selected for inclusion in the sample)

2019 Submission

Testing to identify statistically significant and meaningful differences in performance was conducted using 2017 MIPS performance from registry data provided from CMS. Practices were identified by unique number of TINs, and the 2017 data was from 73 unique TINs. 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 during the 2017 program year. We were unable to determine from our rolled-up data sample the number of clinicians who reported to MIPS as an individual or a 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

Ninety-six practices reported this measure. Data from 786 patient records were submitted for this measure. QOPI measure analytics at the practice level were generated. Practices with fewer than 5 records were not included in calculations.

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

The NPI-level analysis of the 2017 MIPS data was conducted on 254 denominator-eligible patients. Additional descriptive characteristics of the measured patients are unknown. Eligible patients were included in this data set according to the reporting requirements for the 2017 MIPS program year.

2012 Submission

QOPI measure analytics at the practice level were generated. Practices with fewer than 5 records were not included in calculations. Ninety-six practices reported this measure. Data from 786 patient records were submitted for this measure. QOPI measure analytics at the practice level were generated.

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.

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

The 2019 submission also includes additional testing on statistically significant and meaningful differences in performance conducted using 2017 MIPS performance data.

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

Performance measure score: 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 facility performance and the noise is the total variability in measured performance. Reliability at the level of the specific facility 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 facility performance.

Reliability testing was performed by using a beta-binomial model. The beta-binomial model assumes the facility performance score is a binomial random variable conditional on the facility's true value that comes from the beta distribution. The beta distribution is usually defined by two parameters, alpha and beta. Alpha and beta can be thought of as intermediate calculations to get to the needed variance estimates.

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

2008 IRR study: ASCO engaged the Virginia Quality Health Center to conduct an inter-rater reliability study of the QOPI case report form and measures. Trained, independent nurse abstractors served as the 'gold standard' against which practice abstractions were compared for accuracy. Sampling is described above. The 264 sampled records allowed for reliability analysis at a 95% confidence level with a +/- 3.88% marking of error.

Kappa statistics were used to analyze the reliability of the audit data set compared to the submitted data. Kappa statistics are the commonly accepted standard for determining inter-rater reliability in the healthcare setting (Allison, Calhoun, et al, 2000; Cassidy, Marsh, et al, 2002). The Kappa statistic is conceptually similar to the rate of agreement between two reviewers, but it imposes a more stringent standard than simple agreement and mismatch rates. The following standards were used (Cohen, 1960; Sim and Wright, 2005; Feinstein and Cicchetti, 1990):

- Kappa > .0.75 denotes excellent reliability
- Kappa between 0.40 and 0.75 denotes good reliability
- Kappa less than 0.40 denote marginal reliability

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.

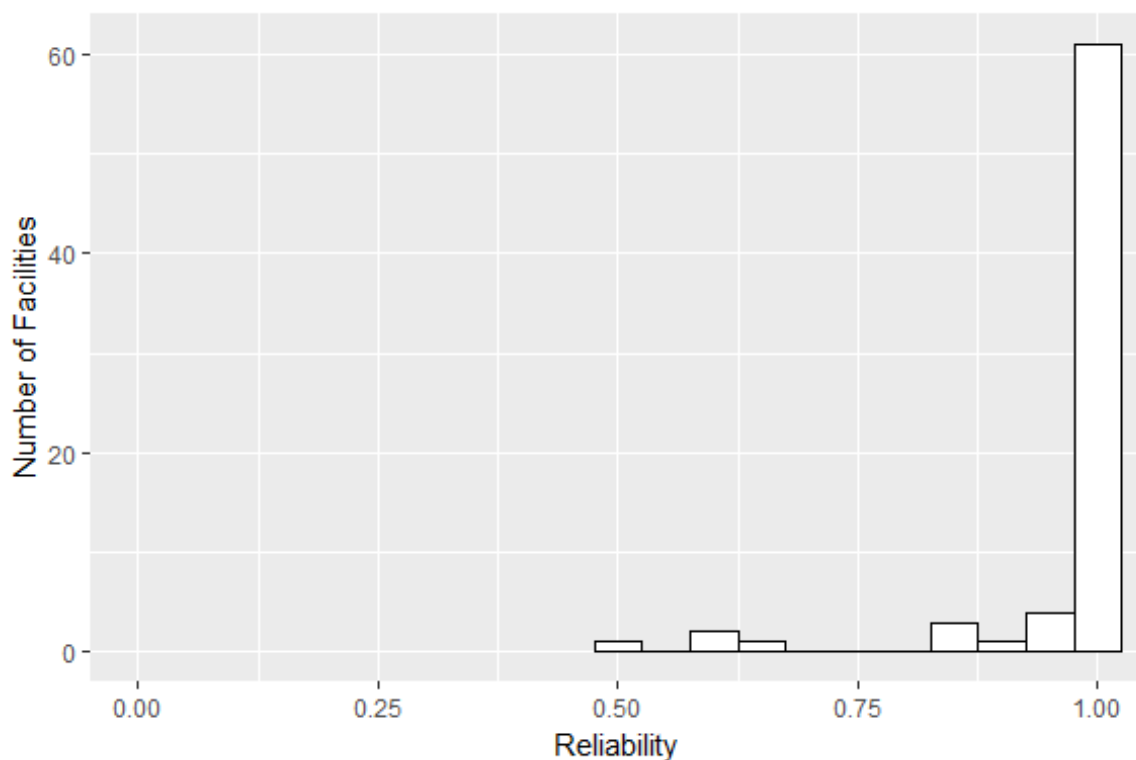
2a2.3. For each level of testing checked above, what were the statistical results from reliability testing? (e.g., percent agreement and kappa for the critical data elements; distribution of reliability statistics from a signal-to-noise analysis)

2019 Submission

Results:

Facility-level Reliability

N	Alpha	Beta	Min	10th Pctl	Median	90th Pctl	Max	Mean
73	1.094	0.1409	0.485	0.8921	1	1	1	0.9657



Interpretation: reliability is excellent. Mean reliability is 97%; the 10th percentile is 89%

2012 Submission

2008 IRR study: A sample of 300 records was planned for re-abstraction in four geographic regions: Midwest, Northeast, South, and West. 50 QOPI practices were randomly selected from the 4 geographic areas and invited to participate. Within each practice, six previously abstracted charts were selected randomly for re-abstraction from the population of 13,561 records submitted in spring 2007 round. Forty-four practices agreed to participate and submitted 264 records (6 per practice).

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.

2008 IRR study: measure level Kappa 0.75 (good reliability). Specifications and instructions were updated based on results.

2010-2011 audit: measure level concordance 96% (valid N=316 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?)

Mean reliability based on signal-to-noise testing was demonstrated to be 97%; the 10th percentile is 89%. Based on these results and in accordance with Adams' definition that reliability between 0.90 – 1.0 is considered very high, ASCO's interpretation is that reliability is excellent for this measure.

ASCO's interpretation of the IRR and concordance analyses from the 2012 submission is that results demonstrated good reliability with a high measure level concordance.

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 is a bivariate analysis that measures the strength of association between two variables and the direction of the relationship. In terms of the strength of relationship, the value of the correlation coefficient varies between +1 and -1. A value of ± 1 indicates a perfect degree of association between the two variables. As the correlation coefficient value goes towards 0, the relationship between the two variables will be weaker. The direction of the relationship is indicated by the sign of the coefficient; a + sign indicates a positive relationship and a – sign indicates a negative relationship.

We hypothesize that there exists a positive association between AJCC stage 1 (T1c) – III HER2 positive breast cancer patients who receive chemotherapy and are administered trastuzumab (NQF #1858/QI 450) and HER2 negative or undocumented breast cancer patients spared treatment with HER2-targeted therapies (NQF #1859/QI 449). ASCO performed a correlation analysis using data from 1858 and measure 1857/QI 449 (HER2 negative or undocumented breast cancer patients spared treatment with HER2-targeted therapies) due to the similarities in patient population and domain. Datasets were reviewed to identify shared providers based on NPI and TIN identifiers. A Pearson correlation analysis was then performed on TINs with denominator counts of ≥ 10 for both measures to evaluate the association between performance scores of these shared providers.

We use the following guidance to describe correlation¹:

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.

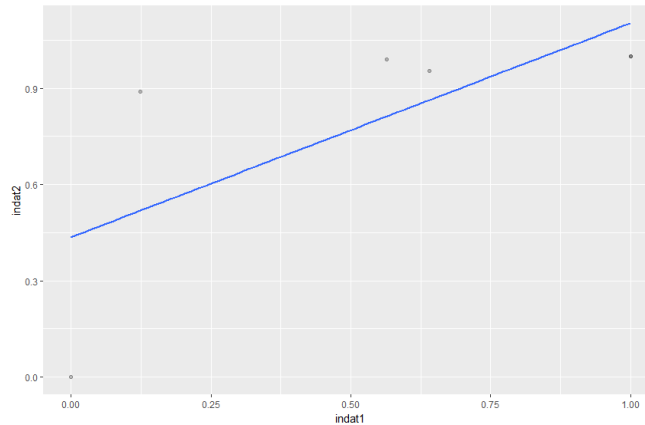
2012 Submission

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

450 with 449: 6 total TINs; correlation = 0.711



2012 Submission

Face validity survey results revealed that 95% 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?)

The results of the analysis of measure QI 450/NQF 1858 and QI 449/NQF 1857 indicate a strong positive correlation. The correlation demonstrates criterion validity of the measure.

Face validity testing demonstrated a vast majority of respondents (95%) 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)

While the rate of exclusions was not provided in our testing data, the previous data element validity analysis provided in the 2012 submission demonstrated a measure level concordance rate of 96% (valid N=316 records), from which we conclude no individual data element (including exceptions) could have had poor validity.

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

N/A

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

N/A

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

If not an intermediate or health outcome, or PRO-PM, or resource use measure, skip to section { [HYPERLINK \l "_bookmark8"](#) }.

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 {HYPERLINK \I “_bookmark7” }

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

The analysis of meaningful differences in performance was analyzed using calculations of 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

Data reported are from the Fall 2011 QOPI round, reflecting data submitted October and November 2011. 96 practices reported this measure. Data from 786 patient records were submitted for this measure.

QOPI measure analytics at the practice level were generated. Practices with fewer than 5 records were not included in calculations.

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

An analysis of 73 unique TINs indicated that 25 percent of TINs have a denominator of two or less, and 50 percent of TINs have a denominator of five or less, and that the measure is heavily skewed with a large proportion of TINs performing perfectly (roughly 55 out of the 73 TINs). Additional details from the TIN-level analysis are provided below.

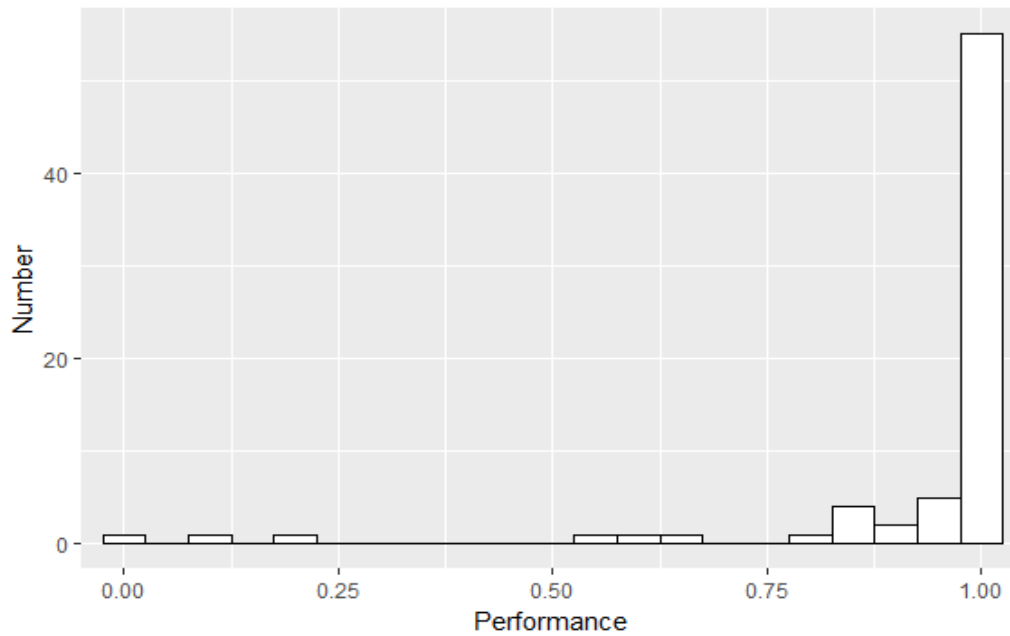
Number of unique entities

Frequency						
73						
Denominators						
Min	Q1	Median	Mean	Q3	Max	Total
1	2	5	20.79	23	206	1518

Measure Distribution

Min	Q1	Median	Mean	Q3	Max	CI.for.mean	Percent.outside.CI
0	0.9853	1	0.9307	1	1	(0.89, 0.98)	90.41

Measure Distribution:



An analysis of 250 unique NPIs indicated results similar to the TIN-level analysis, in that many NPIs have a small denominator, and the majority are already performing at 100 percent. Additional details from the NPI-level analysis are provided below.

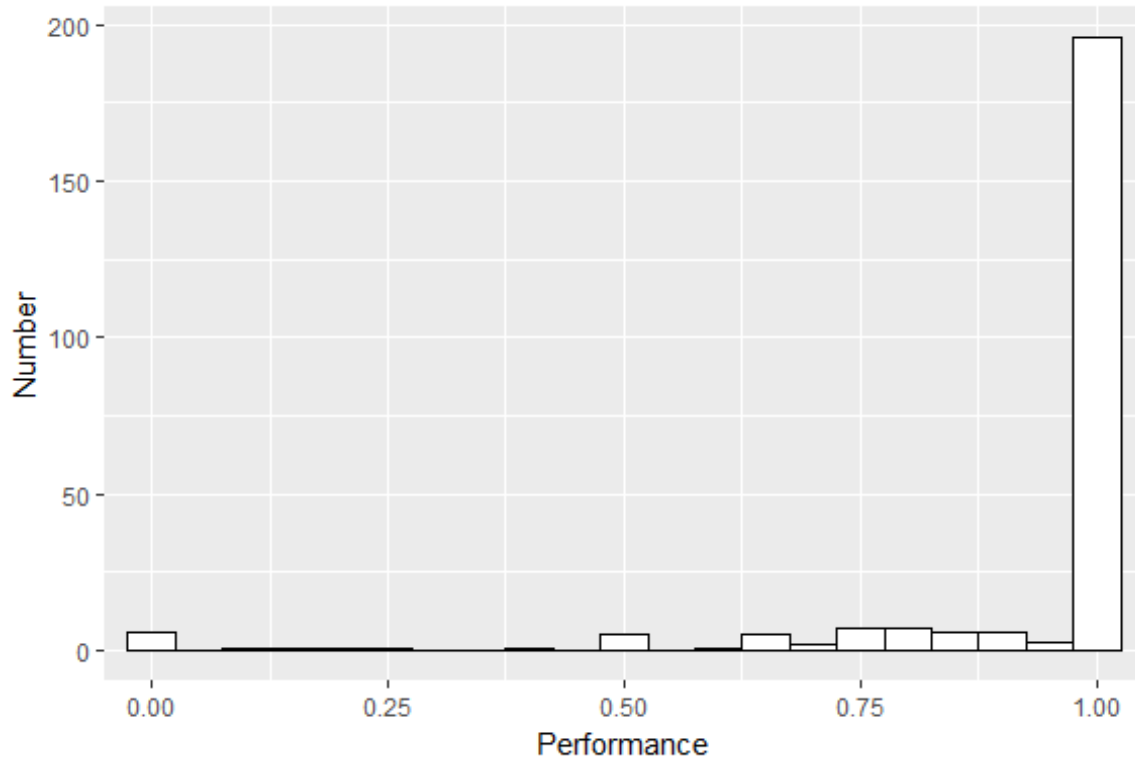
Unique Number of NPIs: 250

Distribution of Measure Denominators and Measure Performance:

Denom:	Min	Q1	Median	Mean	Q3	Max
	0	2	3	6.072	7	45

Measure:	Min	Q1	Median	Mean	Q3	Max	CI.for.mean	Percent.outside.CI
	0	1	1	0.9206	1	1	(0.89, 0.95)	97.19

Measure Distribution:



2012 Submission

For Fall 2011 QOPI round, practice mean = 97%; practice minimum = 60%; practice maximum = 100%

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

Analysis at the TIN level indicated the measure is heavily skewed with a large proportion of TINs performing perfectly (roughly 55 out of the 73 TINs). The majority of TINs perform at 100%, although there are multiple TINs whose performance is below 25%. An analysis of 250 unique NPIs indicated results similar to the TIN-level analysis, in that many NPIs have a small denominator, and the majority are already performing at 100 percent.

Additionally, the 2017 QPP Experience Report Appendix indicates performance on this measure is at 97.51 percent, and 2019 MIPS benchmarking data for QI 450 indicates this measure is topped out. Consistent with these findings, ASCO's interpretation is that NQF 1858 is topped out.

2012 Submission

This measure has been implemented in QOPI for several years. In this self-selected group of oncology practitioners committed to quality assessment and improvement, concordance with this measure has been high overall; however, a proportion of practices (new and experienced) continue to demonstrate sub-optimal variation

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 2017 program year 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 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), Abstracted from a record by someone other than person obtaining original information (e.g., chart abstraction for quality measure or registry)

If other:

3b. Electronic Sources

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

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

Some data elements are in defined fields in electronic sources

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

All the data elements needed for this measure are collected through electronic data or through the use of 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) 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

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.

QOPI® Qualified Clinical Data Registry

This measure has been reported to CMS by the registry as a Qualified Clinical Data Registry. The Quality Oncology Practice Initiative (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.

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 sound 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 in an accountability application 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.

Analysis of MIPS data from 2017 at the TIN level indicated the measure is heavily skewed with a large proportion of TINs performing perfectly (roughly 55 out of the 73 TINs). The majority of TINs perform at 100%, although there are multiple TINs whose performance is below 25%. An analysis of 250 unique NPIs indicated results similar to the TIN-level analysis, in that many NPIs have a small denominator, and the majority are already performing at 100 percent.

Additionally, the 2017 QPP Experience Report Appendix indicates performance on this measure is at 97.51 percent, and 2019 MIPS benchmarking data for QI 450 indicates this measure is topped out. Consistent with these findings, ASCO's interpretation is that NQF 1858 is topped out.

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)

1855 : Quantitative HER2 evaluation by IHC uses the system recommended by the ASCO/CAP guidelines

1857 : HER2 negative or undocumented breast cancer patients spared treatment with HER2-targeted therapies

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 1857 is a complementary measure assessing the inverse of the quality action captured in NQF 1858. Furthermore, because NQF 1857 is endorsed with reserve status and is no longer in use, harmonization is therefore not required. We believe NQF 1855 is a complementary measure assessing HER2 testing, which is an integral component to NQF 1858, and harmonization is not required.

Appendix

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

Attachment:

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, 571-483-1656-

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 breast and gynecologic-oncology Technical Expert Panel (TEP) is a standing multi-disciplinary panel responsible for maintenance and de novo development of ASCO breast and gyn-onc 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:

- Katherine Enright, MD, MPH

Co-Chair - Breast

Cancer Care Ontario – Trillium Health Partners

- Alexi Wright, MD, MPH

Co-Chair – Gyn-onc

Dana-Farber Cancer Institute

- Kerin B. Adelson, MD

Yale School of Medicine Smilow Cancer Center

- Deborah Armstrong, MD

Sidney Kimmel Comprehensive Cancer Center at Johns Hopkins

- Lisa Barbera, MD

Tom Baker Cancer Centre

- Victoria Blinder, MD

Memorial Sloan Kettering Cancer Center 1275 York Ave

- Gary Cohen, MD, FASCO

(Retired) Johns Hopkins School of Medicine

•Neelima Denduluri, MD

US Oncology – Virginia Cancer Specialists

•Nefertiti C. duPont, MD, MPH

North Houston Gynecologic Oncology Surgeons

•Amanda Nickels Fader, MD

Sidney Kimmel Comprehensive Cancer Center at Johns Hopkins

•Carol Hahn, MD, FASTRO

Duke Cancer Center Wake County

•Alexander Melamed, MD, MPH

Columbia University Medical Center

•Monica Morrow, MD, MPH, FASCO, FACS

Memorial Sloan Kettering Cancer Center

•Preeti Sudheendra, MD

MD Anderson Cancer Center at Cooper University Hospital

•William Tew, MD

Memorial Sloan Kettering Cancer Center

•Ann Von Gehr, MD, FACP

Washington Permanente Medical Group

•Jason Wright, MD, FACOG

Columbia University

Measure Developer/Steward Updates and Ongoing Maintenance

Ad.2 Year the measure was first released: 2007

Ad.3 Month and Year of most recent revision: 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.

© 2020 American Medical Association and American Society of Clinical Oncology. All Rights Reserved.

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.

CPT® contained in the Measures specifications is copyright 2004-2019 American Medical Association. LOINC® copyright 2004-2018 Regenstrief Institute, Inc. SNOMED CLINICAL TERMS (SNOMED CT®) copyright 2004-2018 College of American Pathologists. All Rights Reserved.

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

Ad.8 Additional Information/Comments: