**NATIONAL QUALITY FORUM—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:**

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

**Instructions**

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

***Contact NQF staff if more pages are needed.***

* Contact NQF staff regarding questions. Check for resources at [Submitting Standards webpage.](http://www.qualityforum.org/Measuring_Performance/Submitting_Standards.aspx)
* For information on the most updated guidance on how to address social risk factors variables and testing in this form refer to the release notes for version 7.1 of the Measure Testing Attachment.

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

**2a2. Reliability testing** [**10**](#_bookmark0) demonstrates the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise. For **instrument-based measures** (including PRO-PMs) **and composite performance measures**, reliability should be demonstrated for the computed performance score.

**2b1. Validity testing** [**11**](#_bookmark1) demonstrates that the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality. For **instrument- based measures (including PRO-PMs) and composite performance measures**, validity should be demonstrated for the computed performance score.

**2b2. Exclusions** are supported by the clinical evidence and are of sufficient frequency to warrant inclusion in the specifications of the measure; [**12**](#_bookmark2)

# AND

If patient preference (e.g., informed decisionmaking) is a basis for exclusion, there must be evidence that the exclusion impacts performance on the measure; in such cases, the measure must be specified so that the information about patient preference and the effect on the measure is transparent (e.g., numerator category computed separately, denominator exclusion category computed separately). [**13**](#_bookmark3)

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

* **an evidence-based risk-adjustment strategy** (e.g., risk models, risk stratification) is specified; is based on patient factors (including clinical and social risk factors) that influence the measured outcome and are present at start of care;[**14,15**](#_bookmark4) and has demonstrated adequate discrimination and calibration

# OR

* rationale/data support no risk adjustment/ stratification.

**2b4.** Data analysis of computed measure scores demonstrates that methods for scoring and analysis of the specified measure allow for **identification of statistically significant and practically/clinically meaningful** [**16**](#_bookmark5) **differences in performance**;

# OR

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

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

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

**Notes**

1. Reliability testing applies to both the data elements and computed measure score. Examples of reliability testing for data elements include, but are not limited to: inter-rater/abstractor or intra-rater/abstractor studies; internal consistency for multi-item scales; test-retest for survey items. Reliability testing of the measure score addresses precision of measurement (e.g., signal-to-noise).
2. Validity testing applies to both the data elements and computed measure score. Validity testing of data elements typically analyzes agreement with another authoritative source of the same information. Examples of validity testing of the measure score include, but are not limited to: testing hypotheses that the measures scores indicate quality of care, e.g., measure scores are different for groups known to have differences in quality assessed by another valid quality measure or method; correlation of measure scores with another valid indicator of quality for the specific topic; or relationship to conceptually related measures (e.g., scores on process measures to scores on outcome measures). Face validity of the measure score as a quality indicator may be adequate if accomplished through a systematic and transparent process, by identified experts, and explicitly addresses whether performance scores resulting from the measure as specified can be used to distinguish good from poor quality. The degree of consensus and any areas of disagreement must be provided/discussed.
3. Examples of evidence that an exclusion distorts measure results include, but are not limited to: frequency of occurrence, variability of exclusions across providers, and sensitivity analyses with and without the exclusion.
4. Patient preference is not a clinical exception to eligibility and can be influenced by provider interventions.
5. Risk factors that influence outcomes should not be specified as exclusions.
6. With large enough sample sizes, small differences that are statistically significant may or may not be practically or clinically meaningful. The substantive question may be, for example, whether a statistically significant difference of one percentage point in the percentage of patients who received smoking cessation counseling (e.g., 74 percent v. 75 percent) is clinically meaningful; or whether a statistically significant difference of $25 in cost for an episode of care (e.g., $5,000 v.

$5,025) is practically meaningful. Measures with overall less-than-optimal performance may not demonstrate much variability across providers.

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

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

* + 1. **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. **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. **What levels of analysis were tested**? (*testing must be provided for all the levels specified and intended for measure implementation, e.g., individual clinician, hospital, health plan*)

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

* + 1. **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. **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](http://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 that 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.

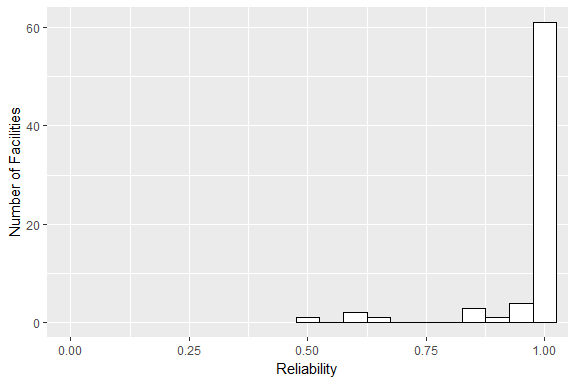
**2a2.3. For each level of testing checked above, what were the statistical results from reliability testing**? (e*.g., percent agreement and kappa for the critical data elements; distribution of reliability statistics from a signal-to-noise analysis*)

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.

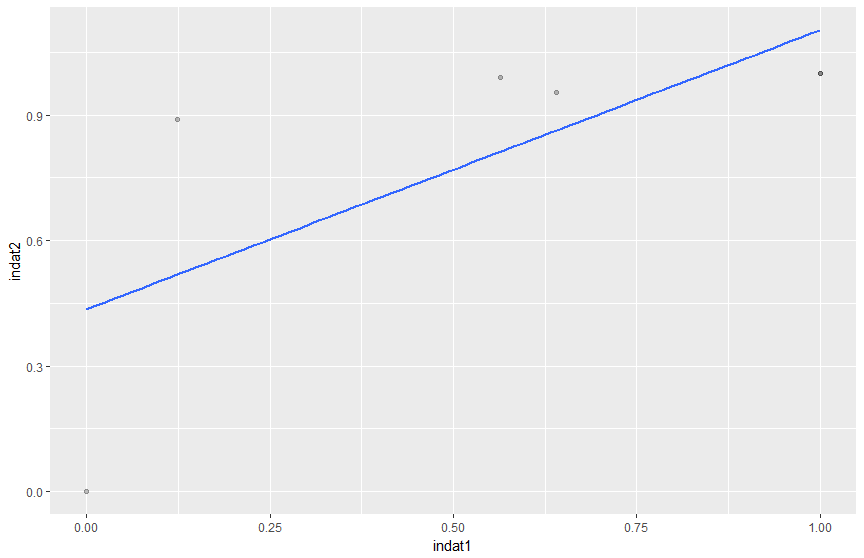
2012 Submission

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 \ I “\_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 |

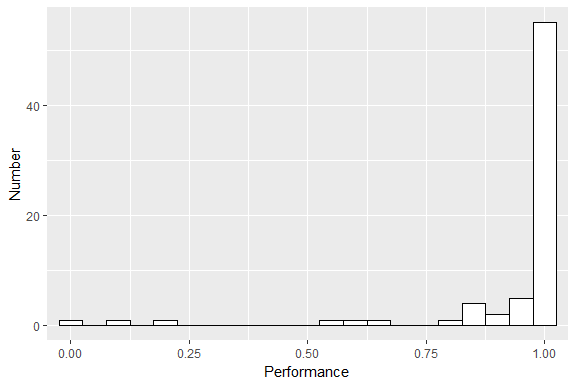
Denominators

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 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:



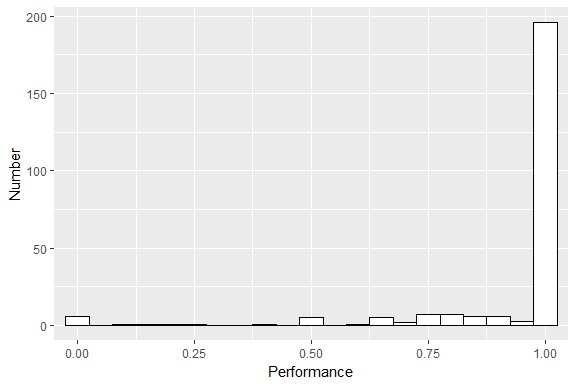
An analysis of 250 unique NPIs indicated results similar to the TIN-level analysis, in that many NPIs have a small denominator, and the majority are already performing at 100 percent. Additional details from the NPI-level analysis are provided below.

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.