**National Quality Forum—Measure Testing (subcriteria 2a2, 2b1-2b6)**

**Measure Number** (*if previously endorsed*)**:** 0384e

**Measure Title**: Oncology: Pain Intensity Quantified – Medical Oncology and Radiation Oncology

**Date of Submission:** 01/05/2020

**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**](#Note10) 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**](#Note11) 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**](#Note12)  **AND**  If patient preference (e.g., informed decision making) 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**](#Note13)  **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**](#Note14)**,**[**15**](#Note15) 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**](#Note16) **differences in performance**;  **OR**  there is evidence of overall less-than-optimal performance.  **2b5.** **If multiple data sources/methods are specified, there is demonstration they produce comparable results**.  **2b6.** Analyses identify the extent and distribution of **missing data** (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias.  **Notes**  **10.** Reliability testing applies to both the data elements and computed measure score. Examples of reliability testing for data elements include, but are not limited to: inter-rater/abstractor or intra-rater/abstractor studies; internal consistency for multi-item scales; test-retest for survey items. Reliability testing of the measure score addresses precision of measurement (e.g., signal-to-noise).  **11.** Validity testing applies to both the data elements and computed measure score. Validity testing of data elements typically analyzes agreement with another authoritative source of the same information. Examples of validity testing of the measure score include, but are not limited to: testing hypotheses that the measures scores indicate quality of care, e.g., measure scores are different for groups known to have differences in quality assessed by another valid quality measure or method; correlation of measure scores with another valid indicator of quality for the specific topic; or relationship to conceptually related measures (e.g., scores on process measures to scores on outcome measures). Face validity of the measure score as a quality indicator may be adequate if accomplished through a systematic and transparent process, by identified experts, and explicitly addresses whether performance scores resulting from the measure as specified can be used to distinguish good from poor quality. The degree of consensus and any areas of disagreement must be provided/discussed.  **12.** Examples of evidence that an exclusion distorts measure results include, but are not limited to: frequency of occurrence, variability of exclusions across providers, and sensitivity analyses with and without the exclusion.  **13.** Patient preference is not a clinical exception to eligibility and can be influenced by provider interventions.  **14.** Risk factors that influence outcomes should not be specified as exclusions.  **15.** With large enough sample sizes, small differences that are statistically significant may or may not be practically or clinically meaningful. The substantive question may be, for example, whether a statistically significant difference of one percentage point in the percentage of patients who received smoking cessation counseling (e.g., 74 percent v. 75 percent) is clinically meaningful; or whether a statistically significant difference of $25 in cost for an episode of care (e.g., $5,000 v. $5,025) is practically meaningful. Measures with overall less-than-optimal performance may not demonstrate much variability across providers. |

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

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

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

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

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

Current Testing

The first dataset used for testing the measure consisted of 2581 qualifying events across 46 oncology physicians. EHR data was pulled from ASCO’s CancerLinQ real world data platform and summarized at the physician level. CancerLinQ collects, organizes, cleans, structures, and analyzes real-world cancer care data from multiple healthcare IT systems and practices across the United States.

The second dataset used for testing consisted of 22,235 qualifying events across 31 providers – both oncology practices and individual oncology physicians. The data was provided by the Center for Medicare & Medicaid Services (CMS) through their PQRS program. PQRS allowed eligible providers (EPs) to submit data directly through a qualified EHR product or through a qualified data submission vendor using certified EHR technology. It also allowed group practices with 2 or more EPs to participate through the group practice reporting option (GPRO) using an EHR direct submission or qualified data submission vendor using certified EHR Technology. PQRS data is summarized at the practice level and includes both EPs participating individually as well as group practices participating through GPRO.

The third dataset used for testing consisted of 1,306 qualifying events on measure PQRS #071: Hormonal Therapy for Stage IC - IIIC Estrogen Receptor/Progesterone Receptor (ER/PR) Positive Breast Cancer across the same 31 providers identified in the second dataset.

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

Previous 2011 Testing

The measurement period (data collected from patients seen) was 1/1/2010 through 12/31/2010. Chart abstraction was performed between 8/8/2011 and 11/3/2011.

Current Testing

EHR data retrieved from CancerLinQ span the time period of 01/01/2018 to 12/31/2018.

Data provided by the Center for Medicare & Medicaid Services (CMS) span the time period from 01/01/2016 to 12/31/2016.

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

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

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

Previous 2011 Testing

Data abstracted from patient records were used to calculate inter-rater reliability for the measure.

862 patient visits were reviewed.

Current Testing

For CancerLinQ dataset, data was extracted and summarized on the physician level for 46 oncology physicians belonging to 3 practices who reported in 2018 on this measure. The practices are a large Midwest integrated health system, a mid-size Maine medical oncology group, and a mid-size California oncology and hematology medical group.

For CMS datasets, data was summarized on the provider level for 31 providers, both individual physicians and groups of physicians, reporting on this measure through the EHR reporting option for CMS’s PQRS in 2016. These 31 providers met the minimum number of quality reporting events (n ≥ 10) for a total of 22,235 quality events (this measure) and 1,306 quality events (PQRS #071). The providers represented all 49 states of the continental Unites States.

**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*)

Previous 2011 Testing

862 patient visits were reviewed for this measure.

Current Testing

CancerLinQ dataset had 2581 cancer patients. All patients qualifying for the measure were included in the dataset. Due to patient privacy concerns and CancerLinQ governing policies no patient specific data was collected.

CMS dataset had 22,235 cancer patients for this measure and 1,306 cancer patients for PQRS #071. All patients qualifying for the measures and associated with a provider meeting the minimum number of 10 qualifying events were included in the dataset. CMS did not provide patient specific characteristics.

**1.7. If there are differences in the data or sample used for different aspects of testing (e.g., reliability, validity, exclusions, risk adjustment), identify how the data or sample are different for each aspect of testing reported below**.

Previous 2011 Testing

Five practice sites representing various types, locations and sizes were identified to participate in testing the PCPI/ASCO/ASTRO-developed measures.

* + Site A: hospital, multi-practice sites in urban, rural and suburban settings; 21 physicians; average 9600 oncology/prostate cancer patient visits per month for MD/NP assessment, chemotherapy; submitted PQRS claims for one measure and utilized a full-fledged EHR.
  + Site B: physician owned private practice, suburban setting; 4 physicians; average 48 oncology/prostate cancer patients seen per day; submitted PQRS claims for one measure and utilized paper medical records.
  + Site C: physician owned private practice, urban setting; 41 physicians; average 2500 oncology/prostate cancer patients seen per month; submitted PQRS claims for two measures and utilized a full-fledged EHR.
  + Site D: academic, suburban setting; 9 physicians; average 240 oncology/prostate cancer patients seen per month; submitted PQRS claims for one measure and utilized paper and EHR.
  + Site E: academic, urban setting; 14 physicians; average 250 oncology/prostate cancer patients seen per month; collected PQRS data on 3 measures and utilized a full-fledged EHR.

Current Testing

For reliability testing of the measure, CancerLinQ dataset was used. For validity testing of the measure, CMS dataset was used. These data sets differ in size, level of analysis, sources of data, physicians/practices, and time periods. Please refer to the comparison table below for more information.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Data Set Name | Time Period | Size (Patients) | Size (Providers) | Data Source | Testing Type | Analysis Level |
| CancerLinQ | CY 2018 | 2581 | 46 | CancerLinQ | Reliability | Physician |
| CMS | CY 2016 | 22,235 | 31 | CMS PQRS Program | Validity | Provider |
| CMS | CY 2016 | 1,306 | 31 | CMS PQRS Program | Validity | Provider |

**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.

Current Testing

Due to patient privacy concerns and CancerLinQ governing policies no patient specific data was collected. Therefore, no social risk factors were available for analysis. Similarly, CMS did not capture nor provide any patient-level socio-demographic (SDS) variables and therefore no social risk factor analysis was conducted.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**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*)

Previous 2011 Testing

Data analysis included:

* Percent agreement; and
* Kappa statistic to adjust for chance agreement.

Current Testing

A split-half reliability psychometric test was chosen to calculate the internal consistency of this measure. A basic assumption of split-half reliability is that the two halves of a physician’s sample of patients should yield similar measure scores and error variances. This comes from another assumption that the measure is focused on the same construct.

Using a random number generator, patients of each of the 46 physicians in the 2018 CancerLinQ dataset were split into two groups. Then, a new measure score was calculated for both group 1 and group 2 of patients for each physician. Once they were calculated, the measures scores of the two groups were correlated and Pearson's r was calculated. Then, the r coefficient was entered into the Spearman-Brown prophecy formula to yield the actual split-half reliability coefficient.

**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*)

Previous 2011 Testing

**N, % Agreement, Kappa (95% Confidence Interval)**

Overall Reliability: 862, **99.9%,** 0.990 (0.970-1.000)

Denominator Reliability: 862, **100.0%,** Kappa is noncalculable\*

Numerator Reliability: 862, **99.9%,** 0.990 (0.970-1.000)

\*Kappa Statistics cannot be calculated because of complete agreement. Confidence intervals cannot be calculated because to do so would involve dividing by zero which cannot be done.

Current Testing

Pearson’s r = 0.7154

Split-half reliability coefficient = 0.8341

**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?*)

Previous 2011 Testing

This measure demonstrates almost perfect reliability, as shown in results from the above analysis.

Current Testing

Overall, correlating two halves of each group of patients for each physician in the CancerLinQ dataset yielded a strong positive relationship. Consequently, we can conclude that this measure is highly reliable.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**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)*

Previous 2011 Testing

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

The expert panel was used to assess face validity of the measure. This panel consisted of 31 members, with representation from the following specialties: oncology, radiation oncology, surgical oncology, urology, gastroenterology, hematology, pathology, colon and rectal surgery, otolaryngology, and pain medicine.

The aforementioned panel was asked to rate their agreement with the following statement:

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

The scale was 1-5, where 1=Strongly Disagree; 3=Neither Disagree nor Agree; 5=Strongly Agree

The expert panel consists of 31 members, whose specialties include oncology, radiation oncology, surgical oncology, urology, gastroenterology, hematology, pathology, colon and rectal surgery, otolaryngology, and pain medicine.

The panel members are as follows:

1. Patricia Ganz, MD (Co-Chair) (Clinical Oncology) University of California – Los Angeles, Los Angeles, CA
2. James Hayman, MD (Co-Chair) (Radiation Oncology) [American Society for Therapeutic Radiology and Oncology (ASTRO](http://www.google.com/url?sa=t&rct=j&q=&esrc=s&frm=1&source=web&cd=2&ved=0CD0QFjAB&url=http%3A%2F%2Fwww.astro.org%2F&ei=fWoHT-DTF8_CgAfW9Z3yCw&usg=AFQjCNF85hwCQ8XSGtxljw3ubhT_a4Wr9Q&sig2=kiRR9P40zvK--HsIVKPnTA)), Ann Arbor MI
3. Joseph Bailes, MD (Clinical Oncology) A Society for Clinical Oncology, The Woodlands, TX
4. Nancy Baxter, MD, PhD (Colorectal Surgery) American Society of Colon and Rectal Surgery Toronto, Ontario Canada
5. Joel V. Brill, MD (Gastroenterology) AGA, Phoenix, AZ
6. Steven B. Clauser, PhD (Outcomes Research) National Cancer Institute, Bethesda, MD
7. Charles Cleeland, PhD (Oncology) McCullough Professor of Cancer Research, Houston, TX
8. J. Thomas Cross, Jr. MD, MPH (Oncology) Colorado Spring, CO
9. Chaitanya R. Divgi, MD (Nuclear Medicine) Professor of Radiology & Chief University of Pennsylvania, Philadelphia, PA
10. Stephen B. Edge, MD (Surgical Oncology) Roswell Park Cancer Institute, Buffalo, NY
11. Patrick L. Fitzgibbons, MD (Oncology) St Jude Medical Center, Fullerton, CA
12. Myron Goldsmith, MD (Oncology) Huntington Beach, CA
13. Joel W. Goldwein, MD (Oncology) IMPAC Medical Systems, Inc., Merion Station, PA
14. Alecia Hathaway, MD, MPH (Oncology) Fort Worth, TX
15. Kevin P. Hubbard, DO (Oncology) Kansas City, MO
16. Nora Janjan, MD, MPSA (Radiation Oncology) University of Texas, Houston, TX
17. Maria Kelly, MB, BCh (Radiation Oncology) ASTRO, Earlysville, VA
18. Wayne Koch, MD (Head and Neck surgery) American Academy of Otolaryngology, Columbia, MD
19. Andre Konski, MD (Radiation Oncology) Fox Chase Cancer Center, Philadelphia, PA
20. Len Lichtenfeld, MD (Oncology) Deputy Chief Medical Officer, American Cancer Society, Atlanta, GA
21. Norman J. Marcus, MD (Anesthesiology and Psychiatry) New York University, New York, NY
22. Catherine Miyamoto, RN, BSN (Oncology) Cancer Center of North Dakota, Grand Forks, ND
23. Michael Neuss, MD (Oncology, Hematology) Oncology Hematology Care, Inc., Cincinnati, OH
24. David F. Penson, MD, MPH (Urology) Associate Professor of Urology and Preventive Medicine, Vanderbilt University Medical Center, Nashville, TN
25. Louis Potters, MD (Radiation Oncology) Chairman of Radiation Medicine, North Shore-NIJ, New Hyde Park, NY
26. John M. Rainey, MD (Medical Oncology) ASCO, Lafayette, LA
27. Christopher M. Rose, MD (Radiation Therapy) Radiation Therapy Center – Beverly Hills, El Segundo, CA
28. Lee Smith, MD (Oncology) Washington Hospital Center, Washington, DC
29. Lawrence A. Solberg, MD, PhD (Oncology) Mayo Clinic, Jacksonville, FL
30. Paul E. Wallner, MD (Radiation Oncology) Willingboro, NJ
31. J. Frank Wilson, MD (Radiation Oncology) Medical College of Wisconsin, Milwaukee, WI

Current Testing

Two CMS datasets were used to perform empirical validity correlation testing of this measure with Breast Cancer: Hormonal Therapy for Stage IC - IIIC Estrogen Receptor/Progesterone Receptor (ER/PR) Positive Breast Cancer (PQRS #071) measure.

Breast Cancer: Hormonal Therapy for Stage IC - IIIC Estrogen Receptor/Progesterone Receptor (ER/PR) Positive Breast Cancer (PQRS #071) was chosen as a suitable candidate for correlation analysis due to both measures’ goal of measuring quality of care for those diagnosed with cancer. We hypothesized that there exists a positive association in performance score between providers that quantify pain intensity for patients with a diagnosis of cancer receiving chemotherapy or radiation therapy (PQRS #143) and those that prescribe tamoxifen or aromatase inhibitor (AI) to female patients with stage I (T1b) through IIIC, ER or PR positive breast cancer (PQRS #071).

Scores for the 31 providers based on 22,235 data points were calculated for the Oncology: Pain Intensity Quantified – Medical Oncology and Radiation Oncology measure. Then, scores for those same 31 providers were calculated based on 1,306 data points for the Breast Cancer: Hormonal Therapy for Stage IC - IIIC Estrogen Receptor/Progesterone Receptor (ER/PR) Positive Breast Cancer measure. Afterwards, the scores were correlated, and Spearman’s rho was calculated for this correlation.

**2b1.3. What were the statistical results from validity testing**? (*e.g., correlation; t-test*)  
Previous 2011 Testing

The results of the expert panel rating of the validity statement were as follows: N = 19; Mean rating = 4.32

Percentage in the top two categories (4 and 5): 84.21%

Frequency Distribution of Ratings

1. 0
2. 1
3. 2
4. 6
5. 10

**N, % Agreement, Kappa (95% Confidence Interval)**

Overall Reliability: 862, **99.9%,** 0.990 (0.970-1.000)

Denominator Reliability: 862, **100.0%,** Kappa is noncalculable\*

Numerator Reliability: 862, **99.9%,** 0.990 (0.970-1.000)

\*Kappa Statistics cannot be calculated because of complete agreement. Confidence intervals cannot be calculated because to do so would involve dividing by zero which cannot be done.

Current Testing

Spearman’s ρ = 0.58

P-value = < 0.001

**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?*)  
Previous 2011 Testing

The measure is valid, as specified.

Current Testing

Oncology: Pain Intensity Quantified – Medical Oncology and Radiation Oncology has a positive correlation with another evidence-based process of care measure (PQRS #071). Therefore, both measures can be utilized to measure the same construct, quality of care for those diagnosed with cancer. Furthermore, this measure has been proven valid using concurrent validity through a comparison with a similar measure.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**2b2. EXCLUSIONS ANALYSIS**

**NA**  **no exclusions — *skip to section*** [***2b3***](#section2b4)

**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*)

Current Testing

Not applicable

**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*)  
Current Testing

Not applicable

**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*)  
Current Testing

Not applicable

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**2b3. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES**  
***If not an intermediate or health outcome, or PRO-PM, or resource use measure, skip to section*** [***2b4***](#section2b5)***.***

**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.**

Current Testing

Not applicable

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

Not applicable

**2b3.3a. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors (clinical factors or social risk factors) used in the statistical risk model or for stratification by risk** (*e.g., potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of p<0.10; correlation of x or higher; patient factors should be present at the start of care*) **Also discuss any “ordering” of risk factor inclusion**; for example, are social risk factors added after all clinical factors?  
Current Testing

Not applicable

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

**Published literature**

**Internal data analysis**

**Other (please describe)**

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

Not applicable

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

Current Testing

Not applicable

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

Not applicable

*Provide the statistical results from testing the approach to controlling for differences in patient characteristics (case mix) below*.  
***If stratified, skip to*** [***2b3.9***](#question2b49)

**2b3.6. Statistical Risk Model Discrimination Statistics** (*e.g., c-statistic, R-squared*)**:**Current Testing

Not applicable

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

Not applicable

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

Not applicable

**2b3.9. Results of Risk Stratification Analysis**:

Current Testing

Not applicable

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

Not applicable

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

Current Testing

Not applicable

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

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

**2b4.1. Describe the method for determining if statistically significant and clinically/practically meaningful differences in performance measure scores among the measured entities can be identified** (*describe the steps―do not just name a method; what statistical analysis was used? Do not just repeat the information provided related to performance gap in 1b)*   
 Previous 2011 Testing

Data analysis performed on the measure included:

Average measure performance rate overall and by site, performance rate range by site and overall standard deviation for the measure.

Current Testing

Measures of central tendency, variability, and dispersion were calculated at the physician level on the CancerLinQ dataset and are displayed in the table below.

|  |  |
| --- | --- |
| N | 46 |
| Minimum | 0.00 |
| 1st Quartile | 95.50 |
| Median | 100.00 |
| 3rd Quartile | 100.00 |
| Maximum | 100.00 |
| Mean | 93.12 |
| Standard Error | 2.0539 |

**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*)  
Previous 2011 Testing

Measure rate without exceptions: N= 862 Mean = 94.0% Standard Deviation= 0.2382

The performance rate by site is as follows, where n is the number of performance events by site:

|  |  |  |
| --- | --- | --- |
| A | 0.9780 | n=183 |
| B | 0.9740 | n=189 |
| C | 0.9730 | n=186 |
| D | 0.9730 | n=188 |
| E | 0.7160 | n=116 |

Current Testing

While most physicians perform in the mid-to-high 90s on this measure, there are still physicians who are failing this measure. Hence, the measure is still capable to meaningfully differentiate among physicians.

**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?*)  
Previous 2011 Testing

The performance rate range is .2620. Although this study captured performance on 862 events, the data were not captured at the physician level, restricting reporting of variation in performance to the organization level only. Additionally, we are unable to present a meaningful calculation of variation in performance across organizations due to the small sample size of sites (n=5) in this study.

Current Testing

Due to the small sample size at the provider level in CancerLinQ dataset, examination of significant differences in the measure performance was kept at the physician level and no provider level statistics were calculated.

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**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*)  
 Current Testing

This test was not performed for this measure.

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

Current Testing

This test was not performed for this measure.

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

Current Testing

This test was not performed for this measure.

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**2b6. MISSING DATA ANALYSIS AND MINIMIZING BIAS**

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

Current Testing

Neither the CancerLinQ dataset nor the two CMS contained missing data so this test was not performed. There is no indication that this missing data was systematic, thus their omission would lead to unbiased performance results.

**2b6.2. What is the overall frequency of missing data, the distribution of missing data across providers, and the results from testing related to missing data?** (*e.g.,**results of sensitivity analysis of the effect of various rules for missing data/nonresponse; if no empirical sensitivity analysis, identify the approaches for handling missing data that were considered and pros and cons of each*)

Current Testing

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

**2b6.3. What is your interpretation of the results in terms of demonstrating that performance results are not biased** due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias**?** (i*.e., what do the results mean in terms of supporting the selected approach for missing data and what are the norms for the test conducted; if no empirical analysis, provide rationale for the selected approach for missing data*)

Current Testing

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