**National Quality Forum**

**Measure missing data in MSF 6.5 from MSF 5.0**

|  |
| --- |
| **NQF #:** 0390 **NQF Project:** Cancer Project |

|  |
| --- |
| **2. RELIABILITY & 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. **(**[**evaluation criteria**](http://www.qualityforum.org/docs/measure_evaluation_criteria.aspx)**)**  Measure testing must demonstrate adequate reliability and validity in order to be recommended for endorsement. Testing may be conducted for data elements and/or the computed measure score. Testing information and results should be entered in the appropriate field. Supplemental materials may be referenced or attached in item 2.1. See [guidance on measure testing](http://www.qualityforum.org/Measuring_Performance/Improving_NQF_Process/Measure_Testing_Task_Force.aspx). |
| **2a2. Reliability Testing.** (*Reliability testing was conducted with appropriate method, scope, and adequate demonstration of reliability*.) |
| **2a2.1 Data/Sample** *(Description of the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)***:**  PCPI Testing Project  Five practice sites representing various types, locations and sizes were identified to participate in testing the 3 PCPI/ASTRO/AUA-developed prostate cancer performance measures.  o 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, chemo; submitted PQRS claims for one measure and utilized a full-fledged EHR.  o 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.  o 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.  o 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.  o 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.  • 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.  **2a2.2 Analytic Method** *(Describe method of reliability testing & rationale)***:**  PCPI Testing Project  Data abstracted from patient records were used to calculate inter-rater reliability for the measure.  91 patient records were reviewed.  Data analysis included:  • Percent agreement; and  • Kappa statistic to adjust for chance agreement.  **2a2.3 Testing Results** *(Reliability statistics, assessment of adequacy in the context of norms for the test conducted)***:**  PCPI Testing Project  N, % Agreement, Kappa (95% Confidence Interval)  Overall Reliability: 91, 98.9%, 0.972 (0.916-1.000)  Denominator Reliability: 91, 100%, Kappa is noncalculable\*  Numerator Reliability: 91, 100%, 0.971 (0.913-1.000)  Exceptions Reliability: 91, 100%, Kappa is noncalculable\*  This measure demonstrates almost perfect reliability, as shown in results from the above analysis.  \*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. |
| **2b. VALIDITY. Validity, Testing, including all Threats to Validity: H M L I** |
| **2b1.1 Describe how the measure specifications** *(measure focus, target population, and exclusions)* **are consistent with the evidence cited in support of the measure focus (***criterion 1c)* **and identify any differences from the evidence:**  The AUA and NCCN guidelines recommend adjuvent hormonal therapy with radiotherapy for high risk prostate cancer patients, for prolonged survival. The measure captures patients receiving external beam radiotherapy in the denominator, and adjuvant hormonal therapy being prescribed in the numerator. Therefore, the evidence directly relates to the specified measure. |
| **2b2. Validity Testing.** (*Validity testing was conducted with appropriate method, scope, and adequate demonstration of validity.)* |
| **2b2.1 Data/Sample** *(Description of the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)***:**  The expert panel consists of 19 members, whose specialties include urology, methodology, clinical oncology, radiation oncology, pathology, family medicine, and consumer and health plan representatives.  The panel members are as follows:  Ian Thompson, MD (Co-Chair, urology)  Steven Clauser, PhD (Co-Chair, methodology)  Peter Albertsen, MD (urology)  Colleen Lawton, MD (radiation oncology)  Charles Bennett, MD, PhD, MPP (clinical oncology)  W. Robert Lee, MD, MS, Med (radiation oncology)  Michael Cookson, MD (urology)  Peter A. S. Johnstone, MD, FACR (radiation oncology)  Gregory W. Cotter, MD (radiation oncology)  David F. Penson, MD, MPH (urology)  Theodore L. DeWeese, MD (radiation oncology)  Stephen Permut, MD (family medicine)  Mario Gonzalez, MD (pathology)  Howard Sandler, MD (radiation oncology)  Louis Kavoussi, MD (urology)  Bill Steirman, MA (consumer representative)  Eric A. Klein, MD (urology)  John T. Wei, MD (urology)  Carol Wilhoit, MD (health plan representative)  **2b2.2 Analytic Method** *(Describe method of validity testing and rationale; if face validity, describe systematic assessment)***:**  All PCPI performance measures are assessed for content validity by 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 (i.e. focus groups) may be convened if there are any remaining concerns related to the content validity of the measures.  Face validity has been quantitatively assessed for this measure. Specifically, the Prostate Cancer Work Group members were asked to empirically assess face validity of the measure. The expert panel consists of 19 members, whose specialties include urology, methodology, clinical oncology, radiation oncology, pathology, family medicine, and consumer and health plan representatives.  Face validity of the measure score as an indicator of quality was systematically assessed as follows:  After the measure was fully specified, the expert panel was asked to rate their agreement with the following statement:  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.  Scale 1-5, where 1=Disagree; 3=Neither Disagree nor Agree; 5=Agree  **2b2.3 Testing Results** *(Statistical results, assessment of adequacy in the context of norms for the test conducted; if face validity, describe results of systematic assessment)***:**  The results of the expert panel rating of the validity statement were as follows: N = 14; Mean rating = 4.57.  Percentage in the top two categories (4 and 5): 92.86%  Frequency Distribution of Ratings  1 – 0  2 – 0  3 – 1  4 – 4  5 – 9 |
| **POTENTIAL THREATS TO VALIDITY**. (*All potential threats to validity were appropriately tested with adequate results.*) |
| **2b3. Measure Exclusions.** (*Exclusions were supported by the clinical evidence in 1c or appropriately tested with results demonstrating the need to specify them.*) |
| **2b3.1 Data/Sample for analysis of exclusions** *(Description of the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)***:**  PCPI Testing Project  • 91 patient records were reviewed for this measure.  • 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.  **2b3.2 Analytic Method** (*Describe type of analysis and rationale for examining exclusions, including exclusion related to patient preference)***:**  Exceptions were analyzed for frequency and variability across providers.  **2b3.3 Results** *(Provide statistical results for analysis of exclusions, e.g., frequency, variability, sensitivity analyses):*  PCPI Testing Project  N, % Agreement, Kappa (95% Confidence Interval)  Exceptions Reliability: 91, 100%, Kappa is noncalculable\*  This measure demonstrates perfect reliability, as shown in results from the above analysis.  The exception rate for this measure is 3.3%  \*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. |
| **2b4. Risk Adjustment Strategy.** (*For outcome measures, adjustment for differences in case mix (severity) across measured entities was appropriately tested with adequate results.*) |
| **2b4.1 Data/Sample** *(Description of the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)***:**  This measure is not risk adjusted.  **2b4.2 Analytic Method (***Describe methods and rationale for development and testing of risk model or risk stratification including selection of factors/variables)***:**  This measure is not risk adjusted.  **2b4.3 Testing Results** *(Statistical risk model: Provide quantitative assessment of relative contribution of model risk factors; risk model performance metrics including cross-validation discrimination and calibration statistics, calibration curve and risk decile plot, and assessment of adequacy in the context of norms for risk models. Risk stratification: Provide quantitative assessment of relationship of risk factors to the outcome and differences in outcomes among the strata)***:**  Not applicable.  **2b4.4 If outcome or resource use measure is not risk adjusted, provide rationale and analyses to justify lack of adjustment:** As a process measure, no risk adjustment is necessary. |
| **2b5. Identification of Meaningful Differences in Performance**. (*The performance measure scores were appropriately analyzed and discriminated meaningful differences in quality.*) |
| **2b5.1 Data/Sample** *(Describe the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)***:**  PCPI Testing Project  • 91 patient records were reviewed for this measure.  • 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.  CMS Physician Quality Reporting Initiative:  Clinical Condition and Measure: #104  2,736 patients were reported on for the 2008 program, the most recent year for which data are available.  In 2009 the following was reported for this measure:  # Eligible Professionals: 4,114  # Professionals Reporting >=1 Valid QDC: 431  % Professionals Reporting >=1 Valid QDC: 10.48%  # Professionals Satisfactorily Reporting: 101  % Professionals Satisfactorily Reporting: 23.43%  **2b5.2 Analytic Method** *(Describe methods and rationale to identify statistically significant and practically/meaningfully differences in performance)***:**  PCPI Testing Project  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.  CMS Physician Quality Reporting Initiative:  The inter-quartile range (IQR) was calculated, which provides a measure of the dispersion of performance.  **2b5.3 Results** *(Provide measure performance results/scores, e.g., distribution by quartile, mean, median, SD, etc.; identification of statistically significant and meaningfully differences in performance)*:  PCPI Testing Project  Measure rate without exceptions: N= 91 Mean = 75.8% Standard Deviation= 0.4305  The performance rate by site is as follows, where n is the number of performance events by site:    A 0.6670 n=21  B 0.8890 n=9  D 0.8000 n=30  E 0.7420 n=31    The performance rate range is .2220. Although this study captured performance on 91 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=4) in this study.  CMS Physician Quality Reporting Initiative  This measure was used in the 2008-2011 CMS Physician Quality Reporting Initiative Claims and Registry options and group reporting option available in 2011.  There is a gap in care as shown by this 2008 data, the only year for which distribution by quartile/decile is available.  83.41% of patients reported on did not meet the measure.  10th percentile: 0.00%  25th percentile: 0.00%  50th percentile: 7.69%  75th percentile: 22.22%  90th percentile: 50.00%  The inter-quartile range (IQR) provides a measure of the dispersion of performance. The IQR is 22.22, and indicates that 50% of physicians have performance on this measure ranging from 0.00% and 22.22%. A quarter of reporting physicians have performance on this measure which is greater than 22.22%, while a quarter have performance on this measure at 0.00%. |
| **2b6.** **Comparability of Multiple Data Sources/Methods.** (*If specified for more than one data source, the various approaches result in comparable scores.*) |
| **2b6.1 Data/Sample** *(Describe the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)***:**  PCPI Testing Project  • 39 Medicare patient records of the 91 patient records were reviewed.  • 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.  **2b6.2 Analytic Method** *(Describe methods and rationale for testing comparability of scores produced by the different data sources specified in the measure)***:**  PCPI Testing Project  Parallel forms reliability testing was performed. PQRS claims were reviewed and compared to a manual review of claims information.  Data analysis included:  • Percent agreement  **2b6.3 Testing Results** *(Provide statistical results, e.g., correlation statistics, comparison of rankings; assessment of adequacy in the context of norms for the test conducted)***:**  PCPI Testing Project  N, % Agreement  39, 100% |
| **2c.** **Disparities in Care:**  **H M L I**  **NA** (*If applicable, the measure specifications allow identification of disparities.*) |
| **2c.1 If measure is stratified for disparities, provide stratified results** *(Scores by stratified categories/cohorts)***:** We encourage the results of this measure to be stratified by race, ethnicity, gender, and primary language, and have included these variables as recommended data elements to be collected.    **2c.2 If disparities have been reported/identified (e.g., in 1b), but measure is not specified to detect disparities, please explain:**  The PCPI advocates that performance measure data should, where possible, be stratified by race, ethnicity, and primary language to assess disparities and initiate subsequent quality improvement activities addressing identified disparities, consistent with recent national efforts to standardize the collection of race and ethnicity data. A 2008 NQF report endorsed 45 practices including stratification by the aforementioned variables.(1) A 2009 IOM report “recommends collection of the existing Office of Management and Budget (OMB) race and Hispanic ethnicity categories as well as more fine-grained categories of ethnicity(referred to as granular ethnicity and based on one’s ancestry) and language need (a rating of spoken English language proficiency of less than very well and one’s preferred language for health-related encounters).”(2)  References:  (1)National Quality Forum Issue Brief (No.10). Closing the Disparities Gap in Healthcare Quality with Performance Measurement and Public Reporting. Washington, DC: NQF, August 2008.  (2)Race, Ethnicity, and Language Data: Standardization for Health Care Quality Improvement. March 2010. AHRQ Publication No. 10-0058-EF. Agency for Healthcare Research and Quality, Rockville, MD. Available at:  http://www.ahrq.gov/research/iomracereport. Accessed May 25, 2010. |
| **2.1-2.3 Supplemental Testing Methodology Information:** |
| **Steering Committee: Overall, was the criterion, *Scientific Acceptability of Measure Properties*, met?**  **(*Reliability and Validity must be rated moderate or high*) Ye**s **No**  **Provide rationale based on specific subcriteria:** |
| **If the Committee votes No, STOP** |

**New reliability testing, empirical validity testing of the measure score, face validity.**

**National Quality Forum—Measure Testing (subcriteria 2a2, 2b2-2b7)**

**Measure Number** (*if previously endorsed*)**:** 390

**Measure Title**: Adjuvant Hormonal Therapy for High Risk Prostate Cancer Patients

**Date of Submission**: 3/11/2016

**Type of Measure:**

|  |  |
| --- | --- |
| Composite – ***STOP – use composite testing form*** | Outcome (*including PRO-PM*) |
| Cost/resource | Process |
| 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, 2b2, 2b3, and 2b5 must be completed.** * **For outcome and resource use measures**, section **2b4** also must be completed. * If specified for **multiple data sources/sets of specificaitons** (e.g., claims and EHRs), section **2b6** 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 (2b2-2b6) 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 20 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 sociodemographic variables and testing in this form refer to the release notes for version 6.6 of the Measure Testing Attachment. |

|  |
| --- |
| **Note: The information provided in this form is intended to aid the Steering 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 **PRO-PMs and composite performance measures**, reliability should be demonstrated for the computed performance score.  **2b2.** **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 **PRO-PMs and composite performance measures**, validity should be demonstrated for the computed performance score.    **2b3.** Exclusions are supported by the clinical evidence; otherwise, they are supported by evidence of sufficient frequency of occurrence so that results are distorted without the exclusion; [**12**](#Note12)  **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**](#Note13)  **2b4.** **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 sociodemographic 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.   **2b5.** 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.  **2b6.** **If multiple data sources/methods are specified, there is demonstration they produce comparable results**.  **2b7.** For **eMeasures, composites, and PRO-PMs** (or other measures susceptible to missing data),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.  **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.23*)** | **Measure Tested with Data From:** |
| abstracted from paper record | abstracted from paper record |
| administrative claims | administrative claims |
| clinical database/registry | clinical database/registry |
| abstracted from electronic health record | abstracted from electronic health record |
| eMeasure (HQMF) implemented in EHRs | eMeasure (HQMF) implemented in EHRs |
| other: Click here to describe | other: Click here to describe |

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

The data source is Registry data from the PQRS program, provided by the Center for Medicare & Medicaid Services (CMS).

**1.3. What are the dates of the data used in testing**? The data are for the time period January 2014 through December 2014 and cover the entire United States.

**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.26*)** | **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*)

The total number of physicians reporting on this measure, via the registry option, in 2014, is 86. Of those, 20 physicians had all the required data elements and met the minimum number of quality reporting events (10) for a total of 634 quality events. For this measure, 23.3 percent of physicians are included in the analysis, and the average number of quality reporting events after exceptions are removed is 21.5 for the remaining 430 events. The range of quality reporting events for 20 physicians included is from 65 to 10. The average number of quality reporting events for the remaining 76.7 percent of physicians that aren’t included is 1.1.

**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*)

There were 430 patients included in this reliability testing and analysis. These were the patients that were associated with physicians who had 10 or more patients eligible for this measure and remained after exceptions were removed.

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

The same data sample was used for reliability testing and exceptions analysis.

1.8 What were the patient-level sociodemographic (SDS) variables that were available and analyzed in the data or sample used? For example, patient-reported data (e.g., income, education, language), proxy variables when SDS data are not collected from each patient (e.g. census tract), or patient community characteristics (e.g. percent vacant housing, crime rate).

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

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

**2a2. RELIABILITY TESTING**

***Note****: If accuracy/correctness (validity) of data elements was empirically tested*, *separate reliability testing of data elements is not required – in 2a2.1 check critical data elements; in 2a2.2 enter “see section 2b2 for validity testing of data elements”; and skip 2a2.3 and 2a2.4.*

**2a2.1. What level of reliability testing was conducted**? (*may be one or both levels*)  
 **Critical data elements used in the measure** (*e.g., inter-abstractor reliability; data element reliability must address ALL critical data elements*)  
 **Performance measure score** (e.g., *signal-to-noise analysis*)  
  
**2a2.2. For each level checked above, describe the method of reliability testing and what it tests** (*describe the steps―do not just name a method; what type of error does it test; what statistical analysis was used*)

Reliability of the computed measure score was measured as the ratio of signal to noise. The signal in this case is the proportion of the variability in measured performance that can be explained by real differences in physician performance. Reliability at the level of the specific physician is given by:

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

Reliability is the ratio of the physician-to-physician variance divided by the sum of the physician-to-physician variance plus the error variance specific to a physician. A reliability of zero implies that all the variability in a measure is attributable to measurement error. A reliability of one implies that all the variability is attributable to real differences in physician performance.

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

Reliability is estimated at two different points, at the minimum number of quality reporting events for the measure and at the mean number of quality reporting events per physician.

**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*)

This measure has 0.73 reliability when evaluated at the minimum level of quality reporting events and 0.85 reliability at the average number of quality events.

**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?*)

Reliability at the minimum level of quality reporting events is high. Reliability at the average number of quality events is high.

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

**2b2. VALIDITY TESTING**

**2b2.1. 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*)

**2b2.2. 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)*

Face validity of the measure score as an indicator of quality was systematically assessed as follows.

After the measure was fully specified, the expert panel was asked to rate their agreement with the following statement:

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.

Scale 1-5, where 1= Strongly Disagree; 3= Neither Agree nor Disagree; 4= Agree; 5= Strongly Agree

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

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

* NQF ICD-10-CM Requirement 1: Statement of intent related to ICD-10 CM

Goal was to convert this measure to a new code set, fully consistent with the original intent of the measure.

* NQF ICD-10-CM Requirement 2: Coding Table

See attachment in S.2b

* NQF ICD-10-CM Requirement 3: Description of the process used to identify ICD-10 codes

The PCPI’s ICD-10 conversion approach was used to identify ICD-10 codes for this measure. The PCPI uses the General Equivalence Mappings (GEMs) as a first step in the identification of ICD-10 codes. We then review the ICD-10 codes to confirm their inclusion in the measure is consistent with the measure intent, making additions or deletions as needed. We have two RHIA-credentialed professionals on our staff who review all ICD-10 coding. For measures included in PQRS, the ICD-10 codes have also been reviewed and vetted by the CMS contractor.  Comments received from stakeholders related to ICD-10 coding are first reviewed internally. Depending on the nature of the comment received, we also engage clinical experts to advise us as to whether a change to the specifications is warranted.

**2b2.3. What were the statistical results from validity testing**? (*e.g., correlation; t-test*)

The expert panel included 21 members. Panel members were comprised of experts from the AUA Quality Improvement and Patient Safety Committee. The list of expert panel members is as follows:

Christopher Tessier, MD (Chair)

Timothy Averch, MD (Vice Chair)

Daniel Barocas, MD

Kristin Chrouser, MD, MPH

Machele Donat, MD

John L. Gore, MD

Fernando Kim, MD

Danil V. Makarov, MD

Jodi Maranchie, MD

Matthew Nielsen, MD

Caleb Nelson, MD

Elliot Paul, MD

John Stoffel, MD

J. Quentin Clemens, MD, MSCI

Roger Dmochowski, MD

Deborah Lightner, MD

Christopher Saigal, MD

J.Stuart Wolf, Jr., MD

**2b2.4. What is your interpretation of the results in terms of demonstrating validity**? (i*.e., what do the results mean and what are the norms for the test conducted?*)

The results of the expert panel rating of the validity statement were as follows: N = 15; Mean rating = 4.5 and 100% of respondents either agree or strongly agree that this measure can accurately distinguish good and poor quality.

Frequency Distribution of Ratings

1 – 0 responses (Strongly Disagree)

2 – 0 responses

3 – 0 responses (Neither Agree nor Disagree)

4 – 7 responses

5 – 8 responses (Strongly Agree)

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

**2b3. EXCLUSIONS ANALYSIS**

**NA**  **no exclusions — *skip to section*** [***2b4***](#section2b4)

**2b3.1. Describe the method of testing exclusions and what it tests** (*describe the steps―do not just name a method; what was tested, e.g., whether exclusions affect overall performance scores; what statistical analysis was used*)

Exceptions include:

* Documentation of medical reason(s) for not prescribing adjuvant hormonal therapy (eg, salvage therapy)
* Documentation of patient reason(s) for not prescribing adjuvant hormonal therapy

Exceptions were analyzed for frequency across providers.

**2b3.2. 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*)

Exceptions Analysis:

Amongst the 20 physicians with the minimum (10) number of quality reporting events, there were a total of 204 exceptions reported. The average number of exceptions per physician in this sample is 10.2. The overall exception rate is 32.2%.

**2b3.3. 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*)

Exceptions are used to remove a patient from the denominator of a performance measure when the patient does not receive a therapy or service AND that therapy or service would not be appropriate due to patient-specific reasons.

Without these being removed, the performance rate would not accurately reflect the true performance of each physician, which would result in an increase in performance failures and false negatives.

AUA recommends that physicians document the specific reasons for exception in patients’ medical records for purposes of optimal patient management and audit-readiness. AUA also advocates for the systematic review and analysis of each physician’s exceptions data to identify practice patterns and opportunities for quality improvement.

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

**2b4. 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*** [***2b5***](#section2b5)***.***

**2b4.1. 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

**2b4.2. If an outcome or resource use measure is not risk adjusted or stratified, provide rationale and analyses to demonstrate that controlling for differences in patient characteristics (case mix) is not needed to achieve fair comparisons across measured entities**.

Not applicable

**2b4.3. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors (clinical factors or sociodemographic 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*)

Not applicable

**2b4.4a. What were the statistical results of the analyses used to select risk factors?**

Not applicable

**2b4.4b. Describe the analyses and interpretation resulting in the decision to select SDS 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)**

Not applicable

**2b4.5. Describe the method of testing/analysis used to develop and validate the adequacy of the statistical model or stratification approach** (*describe the steps―do not just name a method; what statistical analysis was used*)

Not applicable

*Provide the statistical results from testing the approach to controlling for differences in patient characteristics (case mix) below*.  
***If stratified, skip to*** [***2b4.9***](#question2b49)

**2b4.6. Statistical Risk Model Discrimination Statistics** (*e.g., c-statistic, R-squared*)**:**

Not applicable

**2b4.7. Statistical Risk Model Calibration Statistics** (*e.g., Hosmer-Lemeshow statistic*):

Not applicable

**2b4.8. Statistical Risk Model Calibration – Risk decile plots or calibration curves**:

Not applicable

**2b4.9. Results of Risk Stratification Analysis**:

Not applicable

**2b4.10. What is your interpretation of the results in terms of demonstrating adequacy of controlling for differences in patient characteristics (case mix)?** (i*.e., what do the results mean and what are the norms for the test conducted*)

Not applicable

**2b4.11.** **Optional Additional Testing for Risk Adjustment** (*not required, but would provide additional support of adequacy of risk model, e.g., testing of risk model in another data set; sensitivity analysis for missing data; other methods that were assessed*)

Not applicable

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

**2b5. IDENTIFICATION OF STATISTICALLY SIGNIFICANT & MEANINGFUL DIFFERENCES IN PERFORMANCE**

**2b5.1. Describe the method for determining if statistically significant and clinically/practically meaningful differences in performance measure scores among the measured entities can be identified** (*describe the steps―do not just name a method; what statistical analysis was used? Do not just repeat the information provided related to performance gap in 1b)*

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

**2b5.2. 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*)

Based on the sample of 20 included physicians, the mean performance rate is 0.93 the median performance rate is 1.00 and the mode is 1.00. The standard deviation is 0.15. The range of the performance rate is 0.57, with a minimum rate of 0.43 and a maximum rate of 1.00. The interquartile range is 0.06 (.94-1.00).

**2b5.3. 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?*)

The range of performance from .43 to 1.00 suggests there’s clinically meaningful variation across physicians’ performance.

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

**2b6. 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 SDS 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 SDS 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.***

**2b6.1. Describe the method of testing conducted to compare performance scores for the same entities across the different data sources/specifications** (*describe the steps―do not just name a method; what statistical analysis was used*)

This test was not performed for this measure.

**2b6.2. What were the statistical results from testing comparability of performance scores for the same entities when using different data sources/specifications?** (*e.g., correlation, rank order*)

This test was not performed for this measure.

**2b6.3. What is your interpretation of the results in terms of the differences in performance measure scores for the same entities across the different data sources/specifications?** (i*.e., what do the results mean and what are the norms for the test conducted*)

This test was not performed for this measure.

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

**2b7. MISSING DATA ANALYSIS AND MINIMIZING BIAS**

**2b7.1. 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*)

Data are not available to complete this testing.

**2b7.2. 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*)

Data are not available to complete this testing.

**2b7.3. 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*)

Data are not available to complete this testing.