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

**Measure Number** (*if previously endorsed*)**:** 0510

**Measure Title**: Exposure time reported for procedures using fluoroscopy

**Date of Submission**: 1/17/2014

**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). |

|  |
| --- |
| **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 that influence the measured outcome (but not factors related to disparities in care or the quality of care) 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.** Risk models should not obscure disparities in care for populations by including factors that are associated with differences/inequalities in care, such as race, socioeconomic status, or gender (e.g., poorer treatment outcomes of African American men with prostate cancer or inequalities in treatment for CVD risk factors between men and women). It is preferable to stratify measures by race and socioeconomic status rather than to adjust out the differences.  **16.** 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: |

**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*).

Testing was conducted on two occasions for this measure. In 2011, testing was conducted by collecting data from three radiology practice sites representing various types, locations and sizes (hospital and office). Medicare patient medical record data was visually inspected. In 2014, testing was conducted using a Medicare PQRS measure report of claims data.

**1.3. What are the dates of the data used in testing**? Testing in 2011 used data from 1/1/2010 to 12/31/2010. Data abstraction was performed in 2011. Testing in 2014 used data from 1/1/2010 to 12/30/12.

**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: |

**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*)

2011 Testing:

Reliability testing:

* Three radiology practice sites representing various types, locations and sizes were identified to participate in testing measures. The number of physicians per site ranged from approximately 10-1,000 physicians.
* Two of the sites were hospital-based radiology group practices and one was a stand-alone radiology group practice. All three sites were located in urban regions.
* Two of the three sites had EHR testing and one used paper medical records. The abstractors performed a visual inspection of the patient medical record on the sites that employed EHRs. All three measures were tested in claims and the two with E.H.R were compared to the visual inspection of the patient medical record to establish. Individual site performance rates were not obtained because of potential limitations of sample size.
* Patient visit volume ranged from 550-1600 patients, per site, per day.

**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*)   
2011 Testing:

Patient visit volume ranged from 550-1600 patients, per site, per day. Sample size for this measure included a total of 99 records. The sampling strategy was based on a statistical formula that determined the number of patients required for each measure. The initial sample included 30 patients with an oversample of 10. The primary data sources for the abstraction were either the Radiology Information System or EHR. Patient data collected included age and Medicare status, as needed by the measure.

2014 Testing:

PQRS reporting rates were analyzed for 166,801 physicians with 10 or more eligible patients. Patients eligible for the measure included those age 18+ who were Medicare patients.

**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 data/samples described in 1.6 were for reliability testing.

Validity testing was done by expert panel described below in 2b21.

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

**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*)

See section 2b2 for validity testing of data elements.

**For 2011 testing:**

Visual inspection of patient medical record was completed by two trained abstractors to determine inter-rater reliability.

Data analysis included:

* Percent agreement
* Kappa statistic to adjust for chance agreement

In statistics for “Overall reliability,” “Numerator reliability,” and “Denominator reliability,” kappa statistics could not be calculated but are given a value of 1.00 because of complete agreement. Confidence intervals could not be calculated because to do so would involve dividing by zero which cannot be done.

**For 2014 testing:**

An assessment of measure reliability applying a reliability coefficient in the form of the signal to noise ratio (SNR) was conducted. In SNR analysis, reliability is the measure of confidence in differentiating performance between physicians or other providers. The signal is the variability in measured performance that can be explained by real differences in physician performance and the noise is the total variability in measured performance. Reliability is then the ratio of the physician-to-physician variance to the sum of the physician-to-physician variance plus the error variance specific to a physician:

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

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 physician performance. A reliability of 0.70 is generally considered a minimum threshold for reliability and 0.80 is considered very good reliability.

The SNR reliability testing was performed using a beta-binomial model. We limited the analysis to providers with at least 10 patients. Limiting the reliability analysis to only those physicians with a minimum number of events reduces the bias introduced by the inclusion of physicians without a significant numbers of events.

Medicare PQRS data was used for extracting the relevant physician level information. Because we had measure data elements from a large and robust sample of physicians, a de-identified measure reliability analysis could be performed.

**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*)  
From 2011 testing:

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

Overall Reliability (99, 100%, 1.00\* (n/a))

Numerator Reliability (99, 100%, 1.00\* (n/a))

Denominator Reliability (99, 100%, 1.00\* (n/a))

\* Kappa statistics cannot be calculated but are given a value of 1.00 because of complete agreement. Confidence intervals cannot be calculated because to do so would involve dividing by zero which cannot be done.

For 2014 testing:

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Summary of PQRS Reliability Score Stats (2010 - 2012)** | | | | | | | | |  |  |
|  | | **for physicians with 10 or more eligible patients** | | | | | |  |  |  |
|  | |  | |  |  |  | |  |  |  |
|  | | **# of Physicians** | | **Mean Reliability** | **95% Confidence Interval** | | |  |  |  |
|  | | **lower** | **higher** | |  |  |  |
| **2010** | | 64,431 | | 0.9930 | 0.9928 | 0.9933 | |  |  |  |
| **2011** | | 54,782 | | 0.9920 | 0.9917 | 0.9922 | |  |  |  |
| **2012** | | 47,588 | | 0.9922 | 0.9920 | 0.9924 | |  |  |  |
|  | |  | |  |  |  | |  |  |  |
| **Calculation of Physician to Physician Variance** | | | | | | | |  |  |  |
|  | **for physicians with 10 or more eligible patients** | | | | | | |  |  |  |
|  |  | |  | |  |  | |  | **Physician to Physician** | |
|  | **# of Physicians** | | **Observed "p"** | | | **Estimated β distn** | | | **Calculated from β distn** | |
|  | **Mean** | | **Variance** | **Alpha** | **Beta** | | **Mean** | **Variance** |
| **2010** | 64,431 | | 0.049223 | | 0.032485 | 0.021692 | 0.418996 | | 0.049223 | 0.032485 |
| **2011** | 54,782 | | 0.078619 | | 0.048816 | 0.038042 | 0.445840 | | 0.078619 | 0.048816 |
| **2012** | 47,588 | | 0.119017 | | 0.072842 | 0.052302 | 0.387147 | | 0.119017 | 0.072842 |

**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?*)  
2011 testing: The measure data elements can be reliably collected across data sources.

2014 testing: The measure has a high degree of reliability and is able to discriminate effectively between high and low quality performers.

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

**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)*  
This performance measure was assessed for content validity by expert Work Group members during the development process. Additional input on the content validity of draft measures was 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 were reviewed by the expert Work Group and the measure was 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.

The expert panel consists of 14 members, whose specialties include neuroradiology, abdominal radiology, musculoskeletal radiology, cardiac/thoracic radiology, breast imaging, general diagnostic radiology, nuclear medicine, informatics, quality, and physics.

David Seidenwurm, MD, FACR (Chair) (Radiology/Neuroradiology) Sacramento, CA

Dorothy Bulas, MD, FACR (Radiology/Pediatric Radiology) Washington, DC

Robert Henkin, MD, FACR (Nuclear Medicine)

Charles Johnson, MD, FACR (Radiology/Abdominal Radiology) Scottsdale, AZ

David Rubin, MD (Radiology, Musculoskeletal Radiology) Saint Louis, MO

Frank Rybicki, MD (Radiology/Cardiac/thoracic Radiology) Boston, MA

Elizabeth Burnside, MD, MPH (Radiology/Breast Imaging) Madison, WI

Matt Hawkins (Radiology Fellow) Cincinnati, OH

Jonathan Kruskal, MBChB, PhD (Radiology/Abdominal Radiology) Newton, MA

Frank Lexa, MD, MBA (Radiology/Neuroradiology) Wynnewood, PA

Paul Nagy, PhD (Informatics, Quality, Physicist) Baltimore, MD

Donald Renfrew, MD (General Diagnostic Radiology) Sturgeon Bay, WI

Bob Pyatt, MD (General Diagnostic Radiology) Chambersburg, PA

Paul Larson, MD (General Diagnostic Radiology) Madison, WI

Face validity has been quantitatively assessed for this measure. Specifically, the American College of Radiology' Quality Metrics Committee members were asked to empirically assess face validity of the measure. The expert panel consists of 14 members, whose specialties include; neuroradiology, abdominal radiology, musculoskeletal radiology, cardiac/thoracic radiology, breast imaging, general diagnostic radiology, nuclear medicine, informatics, quality, and physics.

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

The results of the expert panel rating of the face validity statement were as follows: N = 7; Mean rating =3.57

Frequency Distribution of Ratings

1 - 0 (Strongly Disagree)

2 - 3

3 - 0 (Neither Disagree nor Agree)

4 - 1

5 - 3 (Strongly Agree)

**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 demonstrate high validity; that is, the measure reflects a quality of care provided, and adequately distinguishes between good and poor quality.

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

**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*)

**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*)

**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*)  
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**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)***.***

Not applicable (no risk adjustment necessary).

**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 (no risk adjustment necessary).

**2b4.3. Describe the conceptual/clinical and statistical methods and criteria used to select patient 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 and not related to disparities*)  
Not applicable (no risk adjustment necessary).

**2b4.4. What were the statistical results of the analyses used to select risk factors?**Not applicable (no risk adjustment necessary).

**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*)

This measure is not stratified.

*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 (no risk adjustment necessary).

**2b4.7. Statistical Risk Model Calibration Statistics** (*e.g., Hosmer-Lemeshow statistic*):   
Not applicable (no risk adjustment necessary).

**2b4.8. Statistical Risk Model Calibration – Risk decile plots or calibration curves**:  
Not applicable (no risk adjustment necessary).

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

Not applicable (no risk adjustment necessary).

**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 (no risk adjustment necessary).

**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 (no risk adjustment necessary).

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

**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)*

For 2011 testing:

An overall performance rate was calculated. Performance rate broken down by site was not assessed due to potential sample size limitations.

For 2014 testing:

As described for reliability testing above, an assessment of the signal to noise ratio (SNR) was conducted. Reliability was measured as the ratio of the physician-to-physician variance to the sum of the physician-to-physician variance plus the error variance specific to a physician:

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

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 physician performance. A reliability of 0.70 is generally considered a minimum threshold for reliability and 0.80 is considered very good reliability.

The SNR reliability testing was performed using a beta-binomial model. We limited the analysis to providers with at least 10 patients.

**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*)  
2011 testing:

PCPI Testing Project Results:

Performance Rate: 86.9%; N=99

2014 testing:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Summary of PQRS Reliability Score Stats (2010 - 2012)** | | | | | |
|  | **for physicians with 10 or more eligible patients** | | | |
|  |  |  |  |  |
|  | **# of Physicians** | **Mean Reliability** | **95% Confidence Interval** | |
|  | **lower** | **higher** |
| **2010** | 64,431 | 0.9930 | 0.9928 | 0.9933 |
| **2011** | 54,782 | 0.9920 | 0.9917 | 0.9922 |
| **2012** | 47,588 | 0.9922 | 0.9920 | 0.9924 |

The reliability of the measure was over 0.99 which implies that nearly all the variance observed was explained by physician differences and not by physician-specific error.

**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?*)

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

**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 criterion is directed 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).* ***If comparability is not demonstrated, the different specifications should be submitted as separate measures.***

**2b6.1. Describe the method of testing conducted to demonstrate comparability of 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*)  
N/A

**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*)  
N/A

**2b6.3. What is your interpretation of the results in terms of demonstrating comparability of 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*)  
N/A

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

**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*)  
2011 testing:

Three radiology practice sites representing various types, locations and sizes were identified to participate in testing the measures. The number of physicians per site ranged from approximately 10-1,000 physicians. Two of the sites were hospital-based radiology group practices and one was a stand-alone radiology group practice. All three sites were located in urban regions. Patient visit volume ranged from 550-1600 patients, per site, per day. Sample size included a total of 99 records for this measure. The data collection period was 1/1/2010- 12/31/2010. Data abstraction was performed in 2011.

An overall performance rate was calculated. Performance rate broken down by site was not assessed due to potential sample size limitations.

The testing sites were queried to determine PQRS involvement for the 2010 PQRS program. It was determined that 2 of the 3 testing sites for the imaging measures were submitting information for the 2010 PQRS Program for this measure.

Abstractors conducted a validation of the PQRS claims data for sites submitting PQRS data. The process began with the identification of a random sample of Medicare claims submitted containing Quality Data Codes for PQRS. The PQRS claim was compared to a visual inspection of the patient's medical record, to determine the reliability of the measure in multiple data sources.

2014 testing:

The data used for testing only included records where data elements were accurately and completely submitted. The analysis includes all physicians who reported the measure for at least 10 patients. For physicians who reported the measure, all data elements were available.

**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*)  
2011 testing:

The data abstracted for the measure was compared to PQRIS claims submissions for the studied site(s). There were 32 PQRIS claims reviewed, of which 100% (32/32) were verified.

2014 testing:

No missing data for providers who reported the measure.

**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*)

Minimal to no bias is expressed through the study approaches used in testing this measure.