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

**Measure Number** (*if previously endorsed*)**:** 00416

**Measure Title**: Diabetic Foot & Ankle Care, Ulcer Prevention – Evaluation of Footwear

**Date of Submission**: 12/6/2013

**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: 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*).

**1.3. What are the dates of the data used in testing**? January 1 – December 31, 2011 and October 1, 2011 – May 1, 2012

**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*) All three physician office sites participating in this measure testing project represented urban settings on the East Coast. The practices each had two or more physicians, with physicians actively involved with APMA.

Two trained data abstractors performed on-site chart reviews the weeks of October 1 and November 5, 2012. Testing was performed on paper medical records at one physician office site and in the electronic health record (EHR) environment for two physician office sites. The case samples for chart reviews were randomly selected from eligible patients seen at two of the test sites between January 1 and December 31, 2011. Due to a change in the billing system, one test site requested a change in the chart sample timeframe to October 1, 2011 through May 1, 2012 to allow for accurate identification of eligible patients.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Site Name** | **Demographics** | **Number of Physicians** | **Number of Office Sites** | **Data Source (EHR, Paper)** | **2011 PQRS Diabetic Foot Care Measures submitted** |
| Practice A | Physician-owned, single specialty | 16 | 12 | EHR | 2 |
| Practice B | Physician-owned, multi-specialty | 9 | 1 | Paper | 2 |
| Practice C | Physician-owned, single specialty | 5 | 5 | EHR | 2 |

**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*)

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Assessment of Evaluation of Footwear** | | | | | | | | | |
|  | | | **Agreement is displayed between PQRS and manual abstraction\*** | | | |  | | |
| **N=286** | **PQRS**  **(n)** | **Manual**  **(n)** | **Yes/Yes\*\*\*\***  **(P/M)** | **Yes/No**  **(P/M)** | **No/Yes**  **(P/M)** | **No/No**  **(P/M)** | **Kappa\*\* Rate** | **95% CI** | **Agreement**  **%** |
| ***Denominator*** | 286 | 286 | 286 | 0 | 0 | 0 | n/c\*\*\* | n/c\*\*\* | 100% |
| ***Numerator*** | 278 | 266 | 262 | 16 | 4 | 4 | 0.256 | (0.036, 0.476) | 93.0% |
| ***Exceptions*** | 0 | 0 | 0 | 0 | 0 | 0 | n/c\*\*\* | n/c\*\*\* | 100% |
| ***Performance rate*** | 97.2% | 93.0% |  |  |  |  |  |  |  |

P = PQRS submitted G-code; M = Manual abstraction; N= sample size; n= number of records; n/c = not calculable

\*Legend of agreement documentation:

* Yes/Yes indicates that both the PQRS G-code and the abstracter indicated “Yes” (per definition of G-code) that the patient met the measure component (numerator, denominator, exception (if applicable);
* Yes/No indicates that the PQRS G-code indicated “Yes” (per definition of G-code) that the patient met the measure component, whereas, the abstractor answered “No” that the patient did not meet the measure component;
* No/Yes indicates that the PQRS G-code indicated “No” (per definition of G-code) that the patient did not meet the measure component, whereas, the abstractor answered “Yes” that the patient did meet the measure component; and
* No/No indicates that both the PQRS G-code (per definition of G-code) and the abstractor indicated “No” that the patient did not meet the measure component.

\*\*The Kappa statistic was calculated to measure the agreement between two data sources by considering agreement between data sources beyond that expected by chance. A kappa of 1.0 indicates perfect agreement and a kappa of 0 indicates agreement attributable solely to chance[[1]](#footnote-1); the higher the kappa, the less likely that agreement was by chance.

**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**.

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

**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*) A Feasibility and Reliability Testing Protocol was drafted by Telligen. *Appendix II* *Podiatry Diabetic Foot and Ankle Measures Feasibility and Reliability Testing Protocol,* includes:

* Objectives
* Number of Records Reviewed
* Sampling Method
* Pre-visit Procedures for On-site Data Abstraction
* On-site Visit Procedure
* Validation of PQRS Claims Data

As a component of feasibility testing, a detailed questionnaire was sent to the sites to explore whether electronic capture of all necessary data elements to compute each measure was inherent in the EHR. Information obtained about each data element included whether the data element was located in a discrete field in the EHR and whether that field was in a standard codified format. The information obtained is discussed further in the measure testing data element location table found later in this document. One practice site did not provide questionnaire results as they were not using an EHR during the measurement timeframe.

Reliability testing of the measures required testing site personnel to use the specifications provided by APMA to complete the necessary programming in the EHR to implement the measures in the EHR and evaluate if a performance report on eligible cases for each of the two measures could be produced. If an EHR-generated report could not be provided, a report of submitted Quality Data codes would be required for the audit. A sample of patients was selected by the site utilizing a randomized sampling strategy to ensure a final sample of 100 patients qualified for the denominator and who received care during the measurement period. As part of the measure reliability testing, two Telligen abstractors performed on**-**site visual inspection of the medical record (each reviewed one half of the sample). To determine if the measures could be collected and calculated reliably, the measures and performance rates were manually constructed for each case. Administrative claim information (claims submitted containing Quality Data codes for PQRS) was compared with the information abstracted from the patient record to determine if the information submitted on the Medicare claim matched the documented care.

**Appendix II – Podiatry Diabetic Foot and Ankle Measures Feasibility and Reliability Testing Protocol**

1. **Objective**

To successfully conduct reliability/validity testing on two NQF time limited endorsed measures\* (NQF 0416 and 0417), both of which have been re-tooled as eMeasures to ultimately assist APMA in attaining full measure endorsement.

**II.** **Number of Records Reviewed**

* Minimum number of 100 individual patient records will be reviewed at each of the four practice sites

1. **Sampling Method**

* To arrive at a sample of 100 records per site, we will over-sample for a total of 110 medical records
* The sample will be from the reporting year of 1/1/2011 to 12/31/2011, who also met the criteria listed below.
  + Patients with diabetes mellitus aged > 18 before the encounter or procedure
  + Payment Source of Medicare (site will pull Medicare claims first and fill in the rest with any claim that satisfies the sampling methodology)
  + Claims were submitted using ICD-9 or SNOMED-CT diagnosis codes **and** CPT or SNOMED encounter or procedure codes **and not** ICD-9 or SNOMED bilateral amputee codes (Page 3 Data Abstraction Definitions/Table 1)
* Telligen is providing the following sampling methodology to the practice sites:
  + Identify records for 100 patients using the coding noted in Table 1 whose Social Security number ends in a specific number, i.e., 2 and 4

1. **Pre-visit Procedures for On-site Data Abstraction**
   * Abstractors will send electronic notification to the practice site that includes the following information:
   * Description of sampling methodology
   * Data Element Tables (location of data in the site’s EHR)
   * Confirmation of availability of staff at practice site
   * Statement of approximate length of on-site visit

* Abstractors will phone office contact 1 week after information is sent to the practice site (discuss any sampling problems and discuss dates for the on-site visit)

1. **On-site Visit Procedure**

* Introduction to staff and EHR
* Discuss security issues/logon information with practice site contact
* Practice site will provide a brief tutorial of the EHR
* Abstractors (2) will need to use two of the practice site computers in order to access the practice site’s EHR. The two abstractors will also have laptops loaded with a pre-approved data collection tool in order to perform the data abstraction

**VI Validation of PQRS Claims Data**

* For the practice sites participating in 2011 PQRS, abstractors will conduct a validation of the PQRS claims data for sites submitting PQRS data. The process includes:
  + Identification of a random sample of Medicare claims submitted containing Quality Data Codes for PQRS
  + Obtain a copy of the Medicare claim from the site
  + Compare the information submitted on the Medicare claim with information in the patient record **(same sample of patients)** to determine if the information submitted matches the PQRS Measure Specifications as posted on the CMS website

Confidentiality of data - In the course of this on-site review of records, Telligen personnel will view Personal Health Information (PHI) as they review patient records. Telligen is a Quality Improvement Organization (QIO) that serves as a health oversight agency for the Centers for Medicare and Medicaid Services (CMS) and is therefore authorized to have access to PHI. Moreover, PHI may be disclosed to Telligen without patient authorization under the HIPAA Privacy Rule at 45 CRF \*\*\*164.512(d).

# 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*) Reliability & Performance rates testing results: evaluation of footwear measure

***Table 2. Reliability Testing Results: Evaluation of Footwear Measure***

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Assessment of Evaluation of Footwear** | | | | | | | | | |
|  | | | **Agreement is displayed between PQRS and manual abstraction\*** | | | |  | | |
| **N=286** | **PQRS**  **(n)** | **Manual**  **(n)** | **Yes/Yes\*\*\*\***  **(P/M)** | **Yes/No**  **(P/M)** | **No/Yes**  **(P/M)** | **No/No**  **(P/M)** | **Kappa\*\* Rate** | **95% CI** | **Agreement**  **%** |
| ***Denominator*** | 286 | 286 | 286 | 0 | 0 | 0 | n/c\*\*\* | n/c\*\*\* | 100% |
| ***Numerator*** | 278 | 266 | 262 | 16 | 4 | 4 | 0.256 | (0.036, 0.476) | 93.0% |
| ***Exceptions*** | 0 | 0 | 0 | 0 | 0 | 0 | n/c\*\*\* | n/c\*\*\* | 100% |
| ***Performance rate*** | 97.2% | 93.0% |  |  |  |  |  |  |  |

P = PQRS submitted G-code; M = Manual abstraction; N= sample size; n= number of records; n/c = not calculable

\*Legend of agreement documentation:

* Yes/Yes indicates that both the PQRS G-code and the abstracter indicated “Yes” (per definition of G-code) that the patient met the measure component (numerator, denominator, exception (if applicable);
* Yes/No indicates that the PQRS G-code indicated “Yes” (per definition of G-code) that the patient met the measure component, whereas, the abstractor answered “No” that the patient did not meet the measure component;
* No/Yes indicates that the PQRS G-code indicated “No” (per definition of G-code) that the patient did not meet the measure component, whereas, the abstractor answered “Yes” that the patient did meet the measure component; and
* No/No indicates that both the PQRS G-code (per definition of G-code) and the abstractor indicated “No” that the patient did not meet the measure component.

\*\*The Kappa statistic was calculated to measure the agreement between two data sources by considering agreement between data sources beyond that expected by chance. A kappa of 1.0 indicates perfect agreement and a kappa of 0 indicates agreement attributable solely to chance[[2]](#footnote-2); the higher the kappa, the less likely that agreement was by chance.

\*\*\*In instances where there is 100% agreement, the kappa statistic cannot be calculated.

\*\*\*\*The kappa is significantly reduced if one classification category dominates. In these cases, the YES category dominates, as one data source had zero NO responses, and the other had very few NO responses. In these cases, with such wide confidence intervals, the kappa is not a valid statistic. This is a limitation of the kappa statistic.

**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?*)

Agreement rates were calculated and reported with kappa statistics with 95% confidence intervals to recognize any agreement that could be attributable to chance alone. **Results:** The measures were found to be highly reliable with agreement rates ranging from 93 to 100%.

**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*)

**Validation Against the Gold Standard Reliability**

Parallel-forms reliability testing was performed by comparing manual abstraction of the data elements necessary to construct the measure from the medical records with Physician Quality Reporting System (PQRS) claims submission. Agreement was calculated between the two methods at the level of the numerator, denominator and exception (if applicable).

To validate inclusion in the numerator, the practice sites provided various identification methods. Two practices provided a report of the sampled list of patients per encounter with the PQRS codes submitted. The third site provided instructions on viewing the billing codes per dates or invoice within each patient’s medical record.

**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)*  
See 2a2.2

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

**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?*)  
Validity demonstrated.

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

**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)***.***

**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**.

**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*)

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

**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*)

*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*)**:**

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

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

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

**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*)

**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*)

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

**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)*

**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*)

**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*)

**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*)

**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*)

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

**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*)

**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*)

**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*)

1. Viera AJ, Garrett JM. Understanding Interobserver Agreement: The Kappa Statistic. Fam Med 2005;37(5):360-3 [↑](#footnote-ref-1)
2. Viera AJ, Garrett JM. Understanding Interobserver Agreement: The Kappa Statistic. Fam Med 2005;37(5):360-3 [↑](#footnote-ref-2)