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

**Measure Number** (*if previously endorsed*): **#2603 (New Measure)**

**Measure Title**: **Diabetes Care for People with Serious Mental Illness: Hemoglobin A1c testing**

**Date of Submission**: **7/25/2014**

**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: 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*).

**Not applicable**

**1.3. What are the dates of the data used in testing**? **2011-2012**

**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*)

**RELIABILITY AND MEANINGFUL DIFFERENCES**

**Three health plans were included in testing and analysis. The plans consisted of a Medicaid plan for non-disabled adults with enrollment of approximately 130,000 members, a Special Needs Plan for dual-eligible members (Medicare and Medicaid) with enrollment of approximately 13,000 members, and a Medicaid plan for disabled adults with enrollment of approximately 13,000 members. The plans were geographically dispersed and included plans from the West, Midwest, and the East regions of the US.**

**SYSTEMATIC EVALUATION OF FACE VALIDITY**

**This measure was tested for face validity with an expert panel (n=16), focus groups (n=29), and public comment (n=20).**

**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*)

**The denominator sample was identified using administrative/claims data for individuals enrolled from January 1, 2011 – December 31, 2012. Given the variation across plans in the prevalence of diabetes and the number of eligible members with serious mental illness, we instructed plans to sample different numbers of patients for chart review to gain a final sample of 250 (127 for Dual SNP, 40 for Medicaid Disabled, and 83 for Medicaid Adult). Please note that rules for calculating the original measure specification consider patients whose medical record is not available for review to be eligible and without any documentation, they are considered to fail the numerator requirements.**

**Of the final sample of 250 patients, 114 were male; 107 were between 18-50 years and 143 were older than 50 years; 93 patients had schizophrenia, 77 had bipolar I disorder, and 80 had major depression.**

**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 full sample of 250 patient medical records was used to test meaningful differences in performance. Of these, 69 patient medical records were double-abstracted and used for inter-rater reliability testing.**

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

**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 was tested by assessing inter-rater reliability. Inter-rater reliability assesses whether two abstractors, reviewing the same data from the same data source, agreed on whether the patient met the requirements for the numerator. Inter-rater reliability was calculated based on data collected by two raters on 69 patient medical records randomly selected across the 3 plans. We used the kappa statistic, a measure of agreement adjusted for chance, to quantify agreement.**

**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*)

**Table 1 shows the number of patients evaluated for the agreement, the percent agreement, and the Kappa statistics (with its 95% confidence interval).**

**Table 1. Percentage of Agreement and Kappa Statistic**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **All Plans** | **Number of patients evaluated for agreement** | **% Agreement** | **Kappa** | **95% C.I.** |
| **Hemoglobin A1c Testing** | **69** | **85.5** | **0.65** | **0.46, 0.85** |

**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?*)

**The Kappa results showed substantial inter-rater reliability for the numerator. This indicates that the measure results can be reliably collected and reported on.**

**For reference, the Kappa statistic has the following interpretation (Landis & Koch, 1977):**

**Table 2. Interpretation of Kappa**

|  |  |
| --- | --- |
| **0.00** | **Poor** |
| **0.01 – 0.20** | **Slight** |
| **0.21 – 0.40** | **Fair** |
| **0.41 – 0.60** | **Moderate** |
| **0.61 – 0.80** | **Substantial** |
| **0.81 – 0.99** | **Almost perfect** |

**Citation:**

**Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977; 33:159-174.**

**This measure is adapted from an existing measure (Comprehensive diabetes care: Hemoglobin A1c testing NQF #0057) that had demonstrated high plan-level reliability among health plans; for example, the beta-binomial statistic for assessing signal-to-noise reliability is 91.7% among Medicaid plans reporting to HEDIS.**

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

**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)*

**Critical Data Element Validity Testing**

**Good inter-rater reliability for data elements also supports the data element validity of the measure.  For reliability testing, two reviewers used the same measure specification would and drew similar conclusions from the same “gold-standard” data source (medical record). This testing demonstrated that two independent reviewers looking at the same full medical record had substantial agreement on every data element and the overall performance measure score.  We believe this testing demonstrates not only reliability but also validity, that is to say the accuracy of the measure specification to identify all data elements from the medical record.  The steps for testing inter-rater reliability were described in section 2a2.**

**Systematic Assessment of Face Validity**

**Our field test addressed the face validity of the measure specification by several types of stakeholder input.**

**A multistakeholder technical expert panel of 16 individuals consisting of health plan representatives, behavioral health and quality measurement experts was convened and provided input throughout the measure development process, including review of the field test results and recommendations for final specifications.**

**In addition, four multistakeholder focus groups that included 29 representatives from Medicaid plans, states, integrated care systems, consumers/advocates, and other health care organizations reviewed and commented on the draft specifications and field test results.**

**We also received feedback from a two-week public comment period hosted on NCQA’s online public comment system. The public comment notification was submitted to stakeholders representing consumers, health plans, clinicians, quality measurement and behavioral health experts (n=20).**

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

**Critical Data Element Validity Testing**

**Substantial inter-rater reliability supports the accuracy of the measure specifications**

**Systematic assessment of face validity**

**Participants in all four of the stakeholder focus groups supported moving forward with this measure. Multiple stakeholders commented that the results were consistent with their expectations. Participants in the focus groups and the TEP felt that the measure had credibility because it was already in use for the general population. Out of 18 total comments that were received from public comment on this measure, 14 (78%) supported or supported the measure with modifications.**

**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 testing results suggest that this measure is valid for assessing HbA1c testing among people with serious mental illness and diabetes. The findings from public comment, focus groups and technical expert panel suggest that the adaptation for the serious mental illness population has specifications that can produce valid results. The substantial inter-rater reliability also provides confidence in the validity of the measure specifications.**

**In addition, this measure is based on an existing measure for the general population of people with diabetes (Comprehensive diabetes care: Hemoglobin A1c testing NQF #0057) which is included in the Medicaid Core set. It is also specified for EHR reporting and included in the CMS EHR Incentive (“Meaningful Use”) Program. The only change is that the denominator is a subpopulation of the existing denominator.**

**The existing measure, Comprehensive Diabetes Care: Hemoglobin A1c testing NQF #0057), is also significantly correlated with other measures of chronic disease care, thus providing evidence of construct validity. For example, the correlation between Hemoglobin A1c testing and Controlling High Blood Pressure was 0.59 (p<.0001) among Medicaid plans in 2012 data. Correlations with other diabetes measures are also significant and range from 0.43 to 0.61.**

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

**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*)

**Not applicable.**

**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*)

**Not applicable.**

**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*)

**Not applicable.**

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

**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 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.**

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

**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)*

**Testing results (N = 3 health plans) did not provide sufficient data to conduct proper statistical tests. While the field test results are limited to 3 health plans, the findings suggest meaningful differences in performance are likely to exist across Medicaid plans.**

**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*)

**The testing results of the measure as specified in the table below showed wide variation among plans, with performance rates of the three plans were 15.7%, 60%, and 65.4%. These results suggest that there are meaningful differences in performance across plans.**

**Table 3: Performance Rate (Medical Records Only)**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **All Sites** | **Dual SNP** | **Medicaid Disabled** | **Medicaid Adult** |
| **Denominator (serious mental illness + diabetes)** | **250** | **127** | **40** | **83** |
| **Numerator (HbA1c Testing)** | **120** | **83** | **24** | **13** |
| **Performance rate** | **48.0%** | **65.4%** | **60.0%** | **15.7%** |

**The performance rate of this measure as specified also varied by age, gender, and mental health diagnosis as shown in the table below. Males tended to be less likely to receive HbA1c testing than females (43% versus 52.2%). The field test data showed differences between the two age groups (18-50 years versus >50 years). However, this result appeared to be related to variations in the age of patients across plans, since most of the patients under age 50 were in the Medicaid Adult plan, which had the lowest performance on this measure. Individuals with bipolar I disorder were less likely to have an HbA1c test than individuals with schizophrenia or major depression.**

**Table 4. Performance by Age, Gender, and Diagnosis (Medical Records Only)**

|  |  |
| --- | --- |
|  | **Performance Rate** |
| **Age** |  |
| **18 to 50 years (inclusive)** | **37.4%** |
| **Greater than 50 years** | **55.9%** |
| **Gender** |  |
| **Male** | **43.0%** |
| **Female** | **52.2%** |
| **Diagnosis** |  |
| **Schizophrenia** | **52.7%** |
| **Bipolar I disorder** | **37.7%** |
| **Major depression** | **52.5%** |

**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?*).

**Our field test showed low overall performance (48%) on this measure and high variation (15.7% - 65.4%) across the three field test plans for the serious mental illness population. For comparison, the average performance rate on the existing measure (Comprehensive diabetes care: Hemoglobin A1c testing NQF #0057) in Medicaid plans reporting to NCQA’s HEDIS program in 2012 was 82.9%, with 10th percentile at 75.6% and 90th percentile at 91%. Stakeholders reported that these findings were likely to be representative. Thus, we interpret the results to suggest that meaningful differences in performance exist and that there is substantial opportunity for improvement.**

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

**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*)

**Not applicable.**

**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*)

**Not applicable.**

**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*)

**Not applicable.**

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

**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*)

**Using standard rules of health plan reporting, sampled patients whose medical records are not available for review are considered numerator failures.**

**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*)

**Not applicable.**

**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*)

**Not applicable.**