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

**Measure Number** (*if previously endorsed*)**:** 0521

**Measure Title**: Heart Failure Symptoms Addressed

**Date of Submission**: 12/23/2013

**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: Electronic Clinical Data | other: Electronic Clinical Data |

**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*).

Home Health OASIS-C

**1.3. What are the dates of the data used in testing**? July 1, 2012 to June 30, 2013

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

|  |  |
| --- | --- |
| **Measure Specified to Measure Performance of:**  **(*must be consistent with levels entered in item S.26*)** | **Measure Tested at Level of:** |
| individual clinician | individual clinician |
| group/practice | group/practice |
| hospital/facility/agency | hospital/facility/agency |
| health plan | health plan |
| other: Click here to describe | other: Click here to describe |

**1.5. How many and which measured entities were included in the testing and analysis (by level of analysis and data source)**? (*identify the number and descriptive characteristics of measured entities included in the analysis (e.g., size, location, type); if a sample was used, describe how entities were selected for inclusion in the sample*)   
The measure included Medicare-certified agencies with at least 20 home health quality episodes beginning between July 1, 2012 and June 30, 2013 and meeting the measure denominator criteria. There were 8,822 such agencies (74.5 percent of the 11,849 agencies with at least one quality episode ending during the same time period). Our sample included all quality episodes (1,932,296 in total) from July 1, 2012 to June 30, 2013.

**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 measure included Medicare-certified agencies with at least 20 home health quality episodes beginning between July 1, 2012 and June 30, 2013 and meeting the measure denominator criteria. There were 8,822 such agencies (74.5 percent of the 11,849 agencies with at least one quality episode ending during the same time period). Our sample included all quality episodes (1,932,296 in total) from July 1, 2012 to June 30, 2013.

**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**.

Not applicable; the same data are used across reliability, validity, and exclusions analyses in this section.

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

**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*)

To address the reliability of the performance scores, Acumen measured the extent to which differences in each quality measure were due to actual differences in agency performance versus variation that arises from measurement error. Statistically, reliability depends on performance variation for a measure across agencies, the random variation in performance for a measure within an agency’s panel of attributed beneficiaries, and the number of beneficiaries attributed to the agency. High reliability for a measure suggests that comparisons of relative performance across agencies are likely to be stable over different performance periods, and that the performance of one agency on the quality measure can confidently be distinguished from another. Potential reliability values range from zero to one, where one (highest possible reliability) means that all variation in the measure’s rates is the result of variation in differences in performance across agencies, while zero (lowest possible reliability) means that all variation is a result of measurement error.

Following the approach described by Adams,[[1]](#footnote-1) Acumen fit a beta-binomial model to estimate measure reliability. The beta-binomial model is appropriate because a particular agency’s measure rate follows a binomial distribution (i.e., all measures are pass/fail), and it is reasonable to assume that the agencies’ true measure rates vary and follow a beta distribution. The true measure rates among the agencies vary because of differences in agency styles, for example. It is reasonable to use the beta distribution to fit the true measure rates because it is a flexible distribution on the interval from 0 to 1, can have any mean on the interval, and can be skewed left, right, or U-shaped.

Equation (1), which is based on the beta-binomial model, shows that reliability is dependent on two variance components: the variation across agencies, and variation within agencies. In general, reliability for agencies will be higher when the measure rates across agencies are more heterogeneous (as measured by the agency-to-agency variation). Agencies with larger samples (n) and pass rates (p) nearer to 0 or 1 will have higher levels of reliability because the agency-specific error is reduced (i.e. the estimated agency rates are more precise).

(1)

Acumen also calculated the test-retest reliability using the intraclass correlation coefficient (ICC) to measure between-agency variation and within-agency variation. Home health episodes within each agency were randomly divided into two separate equally-sized groups. Performance rates were obtained for each set within a measure. Then, a measure-level mean and variance were estimated using the paired performance rates, and an ICC statistic was derived. ICC values that approach 1 indicate that the fraction of the total variance due to between-agency variation is high.

**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*)

**Distribution of Beta Binomial Reliability Scores for Agencies with at Least 20 Valid Episodes**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Mean** | **Minimum** | **10th Percentile** | **25th Percentile** | **Median** | **75th Percentile** | **90th Percentile** | **Maximum** |
| 0.84 | 0.21 | 0.59 | 0.76 | 0.90 | 0.97 | 1.00 | 1.00 |

**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?*)  
Using the beta-binomial model, Acumen concluded that the measure reliability was high. The 25th percentile reliability score is **0.76**, which is above the range considered acceptable (0.70 – 0.80) for drawing inferences about home health agencies. The table below summarizes the distribution of reliability scores for agencies with at least 20 valid episodes.

Furthermore, the ICC coefficient is **0.78** for agencies with at least 40 valid episodes (recall that an ICC statistic is derived from paired performance rates), indicating that most of the total variation is due to between-agency variation.

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

**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)*  
Acumen invited the Home Health Quality Measures Technical Expert Panel to assess the measure’s face validity in December 2010. Face validity refers to the extent to which the measure reflects the quality of care for the specific topic and whether the measure focus is the most important aspect of quality for the specific topic.

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

When asked to rate the face validity of this measure in December 2010, the majority of the TEP members (i.e., 5 out of 8) rated the measure as partially or completely meeting the criteria.

**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?*)  
When asked to rate the face validity of this measure in December 2010, the majority of the TEP members (i.e., 5 out of 8) rated the measure as partially or completely meeting the criteria.

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

**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*)  
As noted in the Denominator Exclusion Details, OASIS data are only collected for particular types of patients. The exclusion of patients who are omitted from OASIS data collection (e.g., those who are non-Medicare/Medicaid, under 18, receiving maternity-related or non-skilled services only) is not based on research evidence but because the measure cannot be calculated due to data limitations.

Acumen calculated the frequency of the measure-specific exclusions by exclusion type. The measure excludes (i) home health episodes for which patient does not have a heart failure diagnosis, and (ii) home health episodes that end in death. Exclusion (i) is justified because the intervention would not be clinically appropriate for episodes in which patient does not have heart failure diagnosis. For exclusion (ii), the information needed to calculate this measure is not collected if the home health episode ends in death.

We also analyzed impact of excluding the long-term episodes. When NQF first considered the *Heart Failure Symptoms Addressed* measure in 2008, the NQF reviewers added a long-term episode exclusion to avoid excessive burden to agencies in reviewing records longer than 60 days. Therefore, the version of themeasure that was endorsed in 2011 included only the short-term episodes. However, the data needed to compute the measure are collected once at the end of each care episode, regardless of whether the episode is short or long-term, so including all episodes in the measure does not increase burden. In preparing this measure for the current NQF reevaluation process, we re-considered the impact of the long-term episodes exclusion.

**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*)

***Heart Failure Symptoms Addressed During All Episodes of Care* Measure Exclusions, By Type**

|  |  |
| --- | --- |
| **Exclusion Type** | **# of Quality Episodes Excluded** |
| Patient death at home (missing M2400A value) | 28,919 |
| No heart failure – no diagnosis found and symptoms not addressed | 2,936,811 |
| **Total # of Quality Episodes Excluded** | 2,965,730 |

**Agency Performance and Number of Agencies Meeting Episode Threshold**

**by Episode Type Restriction**

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Episode Type** | **# of Agencies with ≥ 20 Episodes** | **Agency**  **Average** | **Std. Dev.** | **Average # of Valid Episodes per Agency** | **Percentile** | | | | | | |
| **min** | **p10** | **p25** | **p50** | **p75** | **p90** | **max** |
| Short-Term Episodes Only | 7,566 | 92.8% | 7.3% | 192.1 | 0.0% | 84.6% | 90.3% | 94.4% | 97.4% | 100.0% | 100.0% |
| Long-Term Episodes Only | 5,221 | 93.2% | 7.0% | 77.0 | 1.5% | 84.8% | 90.7% | 95.0% | 97.7% | 100.0% | 100.0% |
| All-Episodes | 8,822 | 92.9% | 7.2% | 216.2 | 0.7% | 85.0% | 90.5% | 94.6% | 97.4% | 100.0% | 100.0% |

**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*)  
The exclusions are supported by sufficient frequency of occurrence. The results would be distorted without the exclusions.

Removing the long-term episodes exclusion (i.e., using all episodes of care rather than the short-term episodes only) does *not* distort the results; the mean agency performance stays almost the same (~93.0%) as a result of this change. Additionally, removing the long-term episodes exclusion (i.e., using all episodes of care rather than the short-term episodes only) actually causes the number of agencies eligible for the measure to increase from 7,566 to 8,822 and the average number of valid episodes per agency to increase from 192 to 216. The table below shows the distribution of agency performance and number of agencies meeting the assessment threshold by episode type restriction.

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

**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; this is a process measure.

**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)*   
The distribution of agency rates was analyzed to determine the inter-quartile range and the 90th vs. 10th percentile differences.

**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*)

***Observed Measure Rates for Agencies with At Least 20 Valid Episodes***

| **Specification** | **Mean** | **St. Dev.** | **Min** | **10th** | **25th** | **50th** | **75th** | **90th** | **Max** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Observed Rates | 92.9% | 7.2% | 0.7% | 85.0% | 90.5% | 94.6% | 97.4% | 100.0% | 100.0% |

Inter-quartile range (75th – 25th) = 97.4% – 90.5% = 6.9%

90th – 10th percentile = 100.0% – 85.0% = 15.0%

**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?*)  
An agency at the 75th percentile still has a performance rate that is 6.9 percentage points higher than an agency at the 25th percentile, meaning the poorer quality agency performs considerably worse on the measure than the agency providing better quality care.

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

**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; this measure uses a single data source.

**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*)  
There are minimal issues with missing data because the OASIS submission system rejects assessments with missing values, and the provider must then resubmit the assessment. The system’s skip pattern specifies that M1500 would not be completed for episodes that end in patient death; these episodes are excluded from the calculations.

**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*)  
There are minimal issues with missing data because the OASIS submission system rejects assessments with missing values, and the provider must then resubmit the assessment. The system’s skip pattern specifies that M1500 would not be completed for episodes that end in patient death; these episodes are excluded from the calculations.

**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*)

There are minimal issues with missing data because the OASIS submission system rejects assessments with missing values, and the provider must then resubmit the assessment. The system’s skip pattern specifies that M1500 would not be completed for episodes that end in patient death; these episodes are excluded from the calculations.

1. For more information about reliability testing for performance measurement, as well as the methodology for constructing the reliability score reported on Table 6, see “Reliability of Provider Profiling: A Tutorial” by John Adams, RAND. http://www.rand.org/pubs/technical\_reports/TR653.html [↑](#footnote-ref-1)