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

**Measure Number** (*if previously endorsed*)**:** Click here to enter NQF number

**Measure Title**: Emergency Department Use without Hospital Readmission During the First 30 Days of Home Health

**Date of Submission**: 2/5/2014

**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*).

Medicare claims

**1.3. What are the dates of the data used in testing**? July 1, 2010 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*)   
For the split-half reliability test, the measure developer included Medicare-certified agencies with at least 40 home health stays beginning between July 2010 to June 2013 and meeting the measure denominator criteria. There were 6,360 such agencies representing a total of 2,827,551 home health stays.

The validity measure testing analyses included Medicare-certified agencies with at least 20 home health stays beginning between July 2010 to June 2013 and meeting the measure denominator criteria. There were 7,565 such agencies representing a total of 2,861,855 home health stays.

The exclusions and risk-adjustment analysis includes all home health stays beginning between July 2010 and June 2012.

**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*)   
In terms of Medicare-certified agencies with at least 40 home health stays beginning between July 2010 to June 2013 and meeting the measure denominator criteria, the data represented 2,505,441 patients. The table below identifies these patients by population group.

**Number/Percentage of Patients Represented in HHAs with At Least 40 Stays, By Population Group**

|  |  |  |  |
| --- | --- | --- | --- |
| **Population Group** | | **# of Patients** | **% of Patients** |
|
| **Total** | | 2,505,441 | 100.0 |
| **Gender** | **Female** | 1,477,297 | 59.0 |
| **Male** | 1,028,144 | 41.0 |
| **Race** | **Black** | 260,027 | 10.4 |
| **Hispanic** | 47,408 | 1.9 |
| **White** | 2,124,319 | 84.8 |
| **Other** | 73,687 | 2.9 |
| **Age** | **<65** | 325,099 | 13.0 |
| **65 - 74** | 790,276 | 31.5 |
| **75 - 84** | 850,980 | 34.0 |
| **85+** | 539,086 | 21.5 |
| **Medicaid  Status** | **Yes** | 545,494 | 21.8 |
| **No** | 1,959,947 | 78.2 |
| **Disabled** | **Yes** | 573,721 | 22.9 |
| **No** | 1,931,720 | 77.1 |

In terms of Medicare-certified agencies with at least 20 home health stays beginning between July 2010 to June 2013 and meeting the measure denominator criteria, the data represented 2,535,844 patients. The table below identifies these patients by population group.

**Number/Percentage of Patients Represented in HHAs with At Least 20 Stays, By Population Group**

|  |  |  |  |
| --- | --- | --- | --- |
| **Population Group** | | **# of Patients** | **% of Patients** |
|
| **Total** | | 2,535,844 | 100.0 |
| **Gender** | **Female** | 1,495,489 | 59.0 |
| **Male** | 1,040,355 | 41.0 |
| **Race** | **Black** | 264,443 | 10.4 |
| **Hispanic** | 49,318 | 1.9 |
| **White** | 2,146,799 | 84.7 |
| **Other** | 75,284 | 3.0 |
| **Age** | **<65** | 329,817 | 13.0 |
| **65 - 74** | 799,029 | 31.5 |
| **75 - 84** | 861,019 | 34.0 |
| **85+** | 545,979 | 21.5 |
| **Medicaid  Status** | **Yes** | 556,229 | 21.9 |
| **No** | 1,979,615 | 78.1 |
| **Disabled** | **Yes** | 581,472 | 22.9 |
| **No** | 1,954,372 | 77.1 |

Exclusions and risk adjustment analysis included home health stays beginning between July 2010 to June 2012. Exclusions analysis began with 2,482,662 potentially eligible stays and identified 1,926,543 eligible stays.

The risk-adjustment model was developed using an 80% random sample and tested using all 1,669,802 stays. (Fewer stays were included in the risk model development, as this development began in early 2013 and 2012 data was not complete at that time.)

**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**.

Differences described in previous question.

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

**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*)

As a measure of internal consistency, the measure developer conducted a split-half reliability test using 100 percent of each home health agency’s patients. Stays for each home health agency were randomly divided into two 50 percent samples, and simulations were run on each 50 percent sample to group the agency into either the “Better than Expected”, “Same as Expected”, or “Worse than Expected” category. Finally, the results between the two samples for each home health agency were compared to assess how consistently the home health agency was grouped into either the “Better than Expected”, “Same as Expected”, or “Worse than Expected” category. The measure developer restricted this analysis to home health agencies with at least 40 valid stays, so that each 50 percent sample had at least 20 stays (which is the minimum number of stays required to protect patient confidentiality in public reporting).

**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*)  
For the split-half test, using each home health agency’s 50 percent samples to produce two simulations and groupings, the majority of the home health agencies were grouped into the same performance category. The figures below depict the results of our split-half test; as represented by the numbers and percentages along the diagonal (i.e., upper-left to bottom-right), 3,855 agencies (78 percent) were grouped into the same performance category as a result of the split-half test. Four-hundred and sixty (10 percent) agencies shifted between the “Better than Expected” and “Same as Expected” categories, and 628 agencies (12 percent) shifted between the “Worse than Expected” and “Same as Expected” categories. Only 10 agencies shifted between the “Better than Expected” and “Worse than Expected” categories.

**Split-Half Test Results**



**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 results show a high level of internal consistency. As a result of the split-half test, the vast majority of the home health agencies were grouped into the same performance category. Some home health agencies shifted between the “Same as Expected” category and “Better than Expected” or “Worse than Expected” categories (which is reasonable because the categorization requires statistical confidence). Finally, transitions between the “Better than Expected” and “Worse than Expected” categories are extremely rare, which shows that the categorization method is robust.

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

**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 Elements - Method*

Review of 2010 Medicare CERT Report. Available at: <https://www.cms.gov/CERT/Downloads/Medicare_FFS_2010_CERT_Report.pdf>

As CMS audits a sample of claims for services provided under Medicare Part B, including emergency department use, as part of annual payment error calculations, additional validity testing of measure elements has not been conducted. The annual payment error calculation for 2010 involved a sample of Medicare claims that were then compared to medical records and included 31,766 claims for services provided under Medicare Part B.

*Empirical Validity Testing - Method*

The measure developer assessed the convergent validity of the measure, which refers to the extent to which measures that are designed to assess the same construct are related to each other. To evaluate the convergent validity of the measure, the measure developer compared the mean performance rates of home health agencies in the “better than expected” category on four measures of home health quality derived from OASIS assessments, compared to the performance of agencies who were not identified as “better than expected” (i.e., home health agencies in the “same as expected” plus “worse than expected” categories).

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

For Part B claims reviewed, 12.9% had some type of payment error. Payment errors include no documentation, insufficient documentation, medically unnecessary service, incorrect coding, and other.

*Empirical Validity Testing - Results*

The table below compares the mean performance rates between home health agencies in the “better than expected” category and other home health agencies (i.e., home health agencies in the “same as expected” plus “worse than expected” categories) on four measures of home health quality derived from OASIS assessments.

**Mean Performance Rates on OASIS Assessment Measures**

|  |  |  |  |
| --- | --- | --- | --- |
| **OASIS Assessment Measure** | **Mean Performance Rate** | | |
| **“Better than Expected” Agencies** | **“Worse than Expected” or “Same as Expected” Agencies** | **Percent Point Difference** |
| How often patients got better at taking their drugs correctly by mouth | 64.6% | 58.1% | 6.5% |
| How often the home health team checked patients for pain | 48.9% | 44.3% | 4.6% |
| How often the home health team determined whether patients received a flu shot for the current flu season | 55.5% | 51.2% | 4.3% |
| How often patients got better at bathing | 65.7% | 62.2% | 3.5% |

**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?*)  
*Critical Data Elements - Interpretation*

The majority of Part B claims with some type of payment error didn’t include an error in the place of service. 12.9% represents a high upper bound for the fraction of claims for outpatient ED visits that do not represent true visits.

*Empirical Validity Testing - Interpretation*

On average, “better than expected” home health agencies perform better on the four OASIS assessment measures above compared to agencies in the pooled “worse than expected” or “same as expected” category, which lends evidence to the measure’s validity. The percent point difference ranges from ~4 percent on the “how often patients got better at bathing” measure to ~7 percent on the “how often patients got better at taking their drugs correctly by mouth” measure. It may be that strong performance on the other measures directly reduces rehospitalizations (e.g., patients who receive the flu vaccine are less likely to catch the flu and require hospitalization). It may also be the case that high quality agencies perform well on both the current measure and other OASIS-based measures due to cultural or organization factors.

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

**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)  
The measure developer calculated frequency of exclusions by exclusion type, using data from July 2010 – June 2012. The measure denominator excludes several types of home health stays:

* First, the measure denominator for the *Rehospitalization During the First 30 Days of Home Health* measure excludes the following home health stays that are also excluded from the all-patient claims-based *NQF 0171 Acute Care Hospitalization* measure:
  + Stays for patients who are not continuously enrolled in fee-for-service Medicare during the measure numerator window
  + Stays that begin with a Low-Utilization Payment Adjustment (LUPA). Stays with four or fewer visits to the beneficiary qualify for LUPAs.
  + Stays in which the patient is transferred to another home health agency within a home health payment episode (60 days).
  + Stays in which the patient is not continuously enrolled in Medicare fee-for-service during the previous six months.
* Second, to be consistent with the *Hospital-Wide All-Cause Unplanned Readmission* measure (as of January 2013), the measure denominator excludes stays in which the hospitalization occurring within 5 days of the start of home health care is not a qualifying inpatient stay. Hospitalizations that do not qualify as index hospitalizations include admissions for the medical treatment of cancer, primary psychiatric disease, or rehabilitation care, and admissions ending in patient discharge against medical advice.
* Third, the measure denominator excludes stays in which the patient receives treatment in another setting in the 5 days between hospital discharge and the start of home health.
* Finally, stays with missing payment-episode authorization strings are excluded.

**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*)

**Measure Denominator Exclusion, July 2010 – June 2012**

|  |  |  |  |
| --- | --- | --- | --- |
| **Home Health Stays** | **# of Stays Excluded** | **% of Stays Excluded** | **# of Stays Remaining** |
| A. Total home health stays beginning within 5 days of hospital discharge | N/A | N/A | 2,482,662 |
| B. Home health stays that meet the denominator criteria for all-patient claims-based *NQF 0171 Acute Care Hospitalization* measure | 418,476 | 16.86 | 2,064,186 |
| C. Home health stays from B that meet the denominator criteria for the *Hospital-Wide All-Cause Unplanned Readmission* measure | 113,454 | 5.50 | 1,950,732 |
| D. Home health stays from C that exclude stays in which the patient receives treatment in another setting in the 5 days between hospital discharge and the start of home health | 23,237 | 1.12 | 1,927,495 |
| E. Home health stays with all risk adjustment data available (i.e., stays with missing payment-episode authorization strings are dropped) | 952 | 0.05 | 1,926,543 |

**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 for the *ED Use without Hospital Readmission During the First 30 Days of Home Health* largely follow the exclusions for the all-patient claims-based *NQF 0173 ED Use without Hospitalization* and the *Hospital-Wide All-Cause Unplanned Readmission* measures. The measure developer imposed these exclusions for consistency with other measures and conducted no additional analyses.

Regarding the exclusion for the stays in which the patient receives treatment in another setting in the 5 days between hospital discharge and the start of home health, the measure developer found that 23,237 of stays (~1 percent) are excluded based on this criterion; this exclusion criterion is justified because the health outcomes of patients who had intervening inpatient (which includes care received at inpatient rehabilitation facilities and long-term care hospitals), emergency department, or skilled nursing facility care in the window between the index hospital discharge and the start of home health care may be affected by this care. We compared each home health agency’s observed rate with and without this exclusion and found the two measures to be highly correlated (overall Pearson correlation coefficient [r] = 0.995).

Regarding the exclusion for stays with missing payment-episode authorization strings, the measure developer found that 952 stays (< 0.1 percent) are excluded based on this criterion; this exclusion criterion is justified because these stays do not include all the information needed for risk adjustment.

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

**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 5 categories of beneficiary-level risk factors (404 significant covariates in total)**

**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 measure is risk-adjusted.

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*)  
The measure developer used a multinomial logistic model to account for beneficiary factors that may affect rates of hospitalization but are outside of the home health agency’s control. Because these measures evaluate two different but related outcomes, one multinomial logistic framework models the three disjoint outcomes: no acute care use (no event), emergency department use without hospital readmission, and rehospitalization. A multinomial logistic model allows for the same risk factors to affect the possible outcomes in different ways while also constraining predicted probabilities of all three events to sum to one hundred percent. The risk adjustment model uses six months of claims prior to the start of home health care to obtain information about the beneficiary. The measure developer identified a set of 404 covariates that consisted of statistically significant predictors of acute care rehospitalization or emergency use without hospital readmission. CMS published the risk adjustment model specifications on the Home Health Quality Initiative page in December 2013.

**2b4.4. What were the statistical results of the analyses used to select risk factors?**Risk model details are attached under the “Specifications” tab.

**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*)  
The same multinomial logit model is used to predict both the Rehospitalization During the First 30 Days of Home Health measure and the ED Use without Hospital Readmission During the First 30 Days of Home Health measure. Of the 1,669,802 qualifying home health stays beginning from July 1, 2010 to June 30, 2012, a random 80 percent sample without replacement was chosen to calibrate the multinomial logit model and to estimate marginal effects for model development purposes. The remaining 20 percent of the stays were used to cross-validate the model.

Risk factors included in the model include prior care setting, health status (measured using HCCs, DRGs, and ADLs), demographic information (measured using age-gender interactions), enrollment status (ESRD and disability), and interactions between one set of the health status covariates. To determine which risk factors should be included in the risk adjustment model, a Wald test of joint restrictions was applied to each variable in each of 1,150 bootstrap samples created using simple random sampling, with replacement, of 80 percent of all home health stays.  The Wald test determined the likelihood that the change in either or both outcomes associated with each covariate was statistically different from zero.  The current risk adjustment model includes only covariates that were significant at a level of 0.05 for either outcome in at least 80 percent of bootstrap samples.  This restriction reduces the number of variables included in the current model, thus streamlining the model and avoiding over-fitting.

To evaluate the impact of each risk factor, the marginal effects were calculated.  The marginal effect represents the relative impact of each risk factor on the outcome.  Each risk factor has an associated marginal effect value that can be interpreted as the change in the population value of the measure if all patients in the population had the risk factor but had the observed distribution of all other risk factors.  Goodness of fit statistics were then calculated for the calibrated model and the 20 percent sample was used for cross-validation.

*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*)**:**The predictive power of the risk adjustment model was evaluated using two measures of predictive power on both the development sample and the validation sample. The two measures of predictive power are the c-statistic and the range of predicted probabilities. Evaluating the model’s predictive power on the development sample shows how well the model predicts outcomes in the data on which it was developed, while evaluating the model using the validation sample shows how well the model predicts outcomes outside the data on which it was developed.

A version of the area under the receiver operating curve (AUC) statistic, also known as the c-statistic, was calculated for each individual logit and for the model overall. The c-statistic measures the ability of a risk adjustment model to differentiate between outcomes without resorting to an arbitrary cutoff point. This analysis averages pair-wise comparisons to extend the standard two-class case to the multi-class form.[[1]](#footnote-1) A model that perfectly discriminates between outcomes would have a c-statistic of 1, while a model that has no predictive power would have a c-statistic of 0.5. To calculate c-statistics for binomial outcomes (i.e., acute care rehospitalization vs. no event and ED use without hospital readmission v. no event), the outlying event was omitted and a generalized logistic estimated on the remaining two outcomes using all the risk factors in the model. A generalized logistic model omitting one event leads to the same coefficients as the full multinomial model. The average of the c-statistics for all possible binomial logistic regressions produces the AUC for the full multinomial model.

The c-statistic for the rehospitalization development sample is 0.693, which is identical to the validation sample value of 0.693, showing that the model differentiates between outcomes as well on new data as it does on the development data. For ED use without hospital readmission, the c-statistic for the development sample is 0.643, which is similar to the validation sample value of 0.642. Finally, the total AUC for the model in the development sample is 0.660, which is comparable to the validation sample value of 0.645.[[2]](#footnote-2) The table below presents these values.

**AUC Statistics**

| **AUC Statistic** | **Development Sample** | **Validation Sample** |
| --- | --- | --- |
| Rehospitalization During the First 30 Days of Home Health c-statistic | 0.693 | 0.693 |
| ED Use without Hospital Readmission During the First 30 Days of Home Health c-statistic | 0.643 | 0.642 |
| Total AUC | 0.660 | 0.645 |

To further evaluate the predictive power of the model, the range of differences between the 90th and 10th percentile of predicted probabilities were calculated. In this case, a larger range of predicted values indicates that the model is better at discriminating between beneficiaries at high risk for rehospitalization or ED use without hospital readmission than beneficiaries at low risk. In the development sample for the multinomial logit model, the range of predicted probabilities for rehospitalization was 4.6 percent to 22.7 percent, and the range was identical in the validation sample. In the development sample, the range of predicted probabilities for ED use without hospital readmission was 5.4 percent to 14.6 percent. In the validation sample, the range was 5.4 percent to 14.7 percent. The table below presents these ranges.

**Range of Differences between 90th and 10th Percentile of Predicted Probabilities**

| **Measure** | **Development Sample** | | **Validation Sample** | |
| --- | --- | --- | --- | --- |
| **Minimum (%)** | **Maximum (%)** | **Minimum (%)** | **Maximum (%)** |
| Rehospitalization During the First 30 Days of Home Health | 4.6 | 22.7 | 4.6 | 22.7 |
| ED Use without Hospital Readmission During the First 30 Days of Home Health | 5.4 | 14.6 | 5.4 | 14.7 |

Finally, the measure developer evaluated the extent to which differences in case-mix would lead to differences in observed rates of ED use without hospital readmission. The table below shows the distribution of expected agency rates of rehospitalization, by agency size. The interquartile ranges, by agency size, range from 0.9 percent for large agencies with 1000+ stays to 1.6 percent for small agencies with 20-49 stays.

**Impact of Risk Adjustment on ED Use without Hospital Readmission Rates, By Agency Size**

| **Total Stays** | **# HHAs** | **Mean** | **St. Dev.** | **Min** | **10th** | **25th** | **50th** | **75th** | **90th** | **Max** | **Interquartile Range** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 20-49 | 1655 | 9.4% | 1.3% | 6.0% | 8.0% | 8.6% | 9.3% | 10.2% | 11.2% | 15.4% | 1.6% |
| 50-99 | 1486 | 9.3% | 1.0% | 6.3% | 8.1% | 8.6% | 9.2% | 9.9% | 10.6% | 13.8% | 1.4% |
| 100–199 | 1385 | 9.3% | 1.0% | 5.8% | 8.2% | 8.6% | 9.2% | 9.9% | 10.6% | 14.2% | 1.3% |
| 200 – 399 | 1244 | 9.2% | 1.0% | 6.1% | 8.1% | 8.5% | 9.1% | 9.8% | 10.4% | 13.0% | 1.2% |
| 400 – 999 | 1115 | 9.0% | 0.8% | 6.1% | 8.1% | 8.5% | 9.0% | 9.5% | 10.0% | 13.3% | 1.0% |
| 1000+ | 680 | 8.9% | 0.7% | 6.7% | 8.1% | 8.4% | 8.8% | 9.3% | 9.7% | 11.6% | 0.9% |

**2b4.7. Statistical Risk Model Calibration Statistics** (*e.g., Hosmer-Lemeshow statistic*):   
Over-fitting occurs when a model can describe the relationship between the covariates and the outcome in the development data set but cannot successfully predict the outcome on a new data set.  To compute the over-fitting indices, the coefficients of the model were first estimated using the development sample.  A logistic regression was then estimated on the validation sample with an intercept and the linear predictor for the probability of an event for a given home health stay in the validation sample.  Values of the intercept far from 0 and values of the coefficient far from 1 provide evidence of over-fitting.  Over-fitting indices were computed separately for the multinomial logit model and the hierarchical-multinomial logit model.

Over-fitting indices were computed and showed no indication that the model was over-fit.  The calibration statistic for rehospitalization produced an intercept of -0.006 and a coefficient of 0.995. With t-statistics of 0.456 and 0.585, these values are not significantly different from 0 and 1, respectively, at the 95% confidence level. In our validation sample, the calibration statistic for ED use without hospital readmission produced an intercept of -0.011 and a coefficient of 0.998.  With t-statistics of 0.456 and 0.180, these values are also not significantly different from 0 and 1 at the 95% confidence level.   In other words, there is no evidence that the model is over-fitting the data for either outcome.

*Over-Fitting Indices*

| **Measure** | **Intercept** | | **Coefficient** | |
| --- | --- | --- | --- | --- |
| Value | Statistically different from 0 at 95% confidence? | Value | Statistically different from 1 at 95% confidence? |
| Rehospitalization During the First 30 Days of Home Health | -0.006 | No | 0.995 | No |
| ED Use without Hospital Readmission During the First 30 Days of Home Health | -0.011 | No | 0.998 | No |

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

**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*)  
The specific set of 404 covariates used in the model consisted of statistically significant predictors of rehospitalization or ED use without hospital readmission during the first 30 days of home health. Risk adjustment compresses the distribution of rehospitalization and ED use without hospital readmission rates and decreases their variability. By taking into account beneficiary characteristics that are outside the provider’s control, the model changes some providers’ relative ranks of rates of rehospitalization and ED use without hospital readmission. The model was found to have considerable predictive power both on the data on which it was developed and on new data and was not determined to be over-fit to the development data.

**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*)

None.

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

**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)*   
Using a three-year reporting period, CMS intends to publicly report the performance of Medicare-certified home health agencies (with at least 20 home health stays) on the rehospitalization and ED use without hospital readmission measures under three performance categories: “better than expected,” “same as expected,” and “worse than expected.” The remainder of this section describes how the measure developer assigned each home health agency to a performance category.

The goal of this method is to assign a home health agency to the “Better than Expected” category if the agency’s rate of ED use without hospital readmission is lower than expected based on patient case mix by a statistically significant amount and to assign a home health agency to the “Worse than Expected” category if the agency’s rate of ED use without hospital readmission is higher than expected based on patient case mix by a statistically significant amount. The size of the difference between a home health agency’s observed rate and expected rate that is statistically significant at a specified level (e.g., 5 percent) depends on the number of home health stays eligible for the measure and the case-mix characteristics of the agency’s specific patients.

Based on patient-level predicted rates from the multinomial logistic model, 20,000 simulated distributions of ED use without hospital readmission rates were generated using SAS, and were used to categorize agencies into the three performance categories. We computed the fraction of simulations that resulted in an ED Use without Readmission rate less than are equal to the observed rate. If this fraction was less than .05, the agency was assigned to the “Better than Expected” category. Analogously, we computed the fraction of simulations that resulted in an ED Use without Readmission rate of greater than or equal to the observed rate. If this fraction was less than .05, the agency was assigned to the “Worse than Expected” category. All other agencies were categorized as “Same as Expected.” Using a value of .05 means that the risk of categorizing a truly average or worse than average agency as better than average is less than 5%. Our accompanying technical brief precisely describes the statistical hypothesis test that this method implements.

With the categorical reporting method, consumers may see that most home health agencies in their area are average, but will be informed if a particular agency is outstanding (i.e., better than expected) or sub-standard (i.e., worse than expected). Additionally, consumers would not make false distinctions between agencies when both home health agencies are performing as expected, even if their observed rates are different.

**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 table below shows the number and percentage of home health agencies, by performance category and size, using home health stays beginning in the period from July 1, 2010 to June 30, 2013. Only agencies with at least 20 stays will have results publicly reported. There were 7,565 such agencies, representing a total of 2,861,855 home health stays and 2,535,844 patients.

**Number of HHAs with “Better than Expected”, “Same as Expected”, or “Worse than Expected”**

**Emergency Department Use without Hospital Readmission Rates, by Number of Stays**

| **Number of Stays** | **Better than Expected** | | **Same as Expected** | | **Worse than Expected** | | **Total** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Count** | **% of Total** | **Count** | **% of Total** | **Count** | **% of Total** |
| <20 | 0 | 0.0% | 4,030 | 98.1% | 78 | 1.9% | 4,108 |
| 20-49 | 32 | 1.9% | 1,568 | 94.7% | 55 | 3.3% | 1,655 |
| 50-99 | 60 | 4.0% | 1,343 | 90.4% | 83 | 5.6% | 1,486 |
| 100-199 | 87 | 6.3% | 1,189 | 85.8% | 109 | 7.9% | 1,385 |
| 200-399 | 96 | 7.7% | 1,008 | 81.0% | 140 | 11.3% | 1,244 |
| 400-999 | 113 | 10.1% | 779 | 69.9% | 223 | 20.0% | 1,115 |
| 1000+ | 155 | 22.8% | 387 | 56.9% | 138 | 20.3% | 680 |
| **Total** | **543** | **4.7%** | **10,304** | **88.3%** | **826** | **7.1%** | **11,673** |

**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?*)

With the categorical reporting method, consumers may see that most home health agencies in their area are average (i.e., same as expected, which applies to 88.3 percent of home health agencies), but will be informed if a particular agency is outstanding (i.e., better than expected, which applies to 4.7 percent of home health agencies and 7.2 percent of agencies with 20 or more stays) or sub-standard (i.e., worse than expected, which applies to 7.1 percent of home health agencies and 9.9 percent of agencies with 20 or more stays). Additionally, consumers would not make false distinctions between agencies when both home health agencies are performing as expected, even if their observed rates are different.

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

**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*)  
The measure developer found that 952 stays (< 0.1 percent) had missing payment-episode authorization strings. Because these stays do not include all the information needed for risk adjustment, they are excluded from the measure. Additionally, patients with insufficient Medicare FFS enrollment to allow for measure calculation are excluded from the measure.

**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*)  
The measure developer found that 952 stays (< 0.1 percent) had missing payment-episode authorization strings. Because these stays do not include all the information needed for risk adjustment, they are excluded from the measure. Additionally, patients with insufficient Medicare FFS enrollment to allow for measure calculation are excluded from the measure.

**2b7.3. What is your interpretation of the results in terms of demonstrating that performance results are not biased** due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias**?** (i*.e., what do the results mean in terms of supporting the selected approach for missing data and what are the norms for the test conducted; if no empirical analysis, provide rationale for the selected approach for missing data*)

The measure developer found that 952 stays (< 0.1 percent) had missing payment-episode authorization strings. Because these stays do not include all the information needed for risk adjustment, they are excluded from the measure. Additionally, patients with insufficient Medicare FFS enrollment to allow for measure calculation are excluded from the measure.

1. For more information on this extension of the c-statistic, please refer to: David J. Hand and Robert J. Till, “A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems.” Ed. David W. Aha*. Machine Learning* 45 (2001): 171-186. [↑](#footnote-ref-1)
2. The total area under the curve is an assessment of the overall model fit obtained by averaging the c-statistics for the individual logits, which in this case is the two c-statistics shown as well as the c-statistic between rehospitalization and ED use without hospital readmission, which is not shown. For more information on this statistic, refer to the footnote above. [↑](#footnote-ref-2)