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

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

**Measure Title**: Hospitalizations per 1000 Medicare fee-for-service (FFS) Beneficiaries

**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: Medicare denominator file 2012 | ☒ other: Medicare denominator file 2012 |

**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 Part A claims and the Denominator File

**1.3. What are the dates of the data used in testing**? Click here to enter date range

* Annual Measure - 2009-2012
* Quarterly Measure - Q1 2009 through Q1 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: States/Territories & Communities | ☒ other: States/Territories, Communities, Hospital Referral Regions (HRRs) |

**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*)

53 States/Territories, 327 Communities, and 306 Hospital Referral Regions (HRRs) (based on a set of ZIP Codes).

**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*)

About 40 million Medicare beneficiaries, with info about sex, race/ethnicity, age, dual eligibility (Medicare and Medicaid), and urban/rural status. The table below provides descriptive characteristics of the Medicare FFS beneficiaries for 2012 for states and communities used in the analysis. The community testing data set is a subset (40.1%) of the State testing data set. N represents the number of Medicare FFS beneficiaries.

| Beneficiary Sample Summary (2012) | | | | | |
| --- | --- | --- | --- | --- | --- |
|  |  | State Testing | | Community Testing | |
| Variable | Level | N | Percent | N | Percent |
|  | Total | 39,478,873 | 100.0% | 15,814,412 | 100.0% |
| Sex | Male | 18,142,773 | 46.0% | 7,222,498 | 45.7% |
| Female | 21,336,100 | 54.0% | 8,591,914 | 54.3% |
| Race | White | 32,423,180 | 82.1% | 12,684,504 | 80.2% |
| Black | 4,103,687 | 10.4% | 1,961,727 | 12.4% |
| Hispanic | 956,833 | 2.4% | 409,486 | 2.6% |
| Asian | 795,012 | 2.0% | 289,160 | 1.8% |
| N. American Native | 209,758 | 0.5% | 71,907 | 0.5% |
| Other | 690,043 | 1.7% | 270,078 | 1.7% |
| Unknown | 300,360 | 0.8% | 127,550 | 0.8% |
| Age Group | < 50 | 2,682,147 | 6.8% | 1,130,784 | 7.2% |
| 50-64 | 4,592,161 | 11.6% | 1,884,996 | 11.9% |
| 65-69 | 10,522,952 | 26.7% | 4,224,836 | 26.7% |
| 70-74 | 7,169,724 | 18.2% | 2,826,309 | 17.9% |
| 75-79 | 5,304,110 | 13.4% | 2,084,825 | 13.2% |
| 80-84 | 4,192,821 | 10.6% | 1,658,262 | 10.5% |
| >= 85 | 5,014,958 | 12.7% | 2,004,400 | 12.7% |
| Dual | Medicare only | 31,766,417 | 80.5% | 12,579,768 | 79.5% |
| Medicare/  Medicaid | 7,712,456 | 19.5% | 3,234,644 | 20.5% |
| Urban Type | Urbanized Area | 27,737,539 | 70.3% | 12,404,783 | 78.4% |
| Urban Cluster | 7,199,780 | 18.2% | 2,102,354 | 13.3% |
| Non-Urban Area | 4,507,759 | 11.4% | 1,298,125 | 8.2% |
| UTD | 33,795 | 0.1% | 9,150 | 0.1% |

The 306 HRRs were used for one aspect of testing. The number of FFS beneficiaries and corresponding percents for HRRs would be very close to those for State testing, but not identical since Puerto Rico and US Virgin Islands ZIP Codes are not assigned to any of the 306 HRRs.

Please see the Appendix for community level characteristics.

**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**.

N/A

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

**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*)

Test-retest approach:  We randomly split the approximately 40 million beneficiaries into two separate samples and computed measure rates at the State/Territory level and at the Community level for each sample.  Samples were compared using correlation statistics and quintile 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*)

2012 Annual Measure for States/Territories

| Hospitalizations per 1000 Summary | | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
|  | N | Mean | Std Dev | Median | Minimum | Maximum |
| Sample 1 | 53 | 272.84 | 47.09 | 279.73 | 147.78 | 337.47 |
| Sample 2 | 53 | 273.21 | 46.11 | 276.49 | 147.24 | 337.12 |

| Hospitalizations per 1000 Correlations | |
| --- | --- |
|  | Correlation (p-value) |
| Pearson Correlation | 0.99808 (<.0001) |
| Spearman Correlation (uses ranks) | 0.99653 (<.0001) |

|  | Hospitalizations per 1000 Quintile Comparison | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | Quintiles Sample 2 | | | | | |
|  |  | 1 | 2 | 3 | 4 | 5 | Total |
| Quintiles  Sample 1 | 1 | 11 | 0 | 0 | 0 | 0 | 11 |
| 2 | 0 | 9 | 2 | 0 | 0 | 11 |
| 3 | 0 | 2 | 7 | 1 | 0 | 10 |
| 4 | 0 | 0 | 1 | 10 | 0 | 11 |
| 5 | 0 | 0 | 0 | 0 | 10 | 10 |
|  | Total | 11 | 11 | 10 | 11 | 10 | 53 |

Quintile Agreement: 89% (47/53)

Weighted Kappa: 0.9292 (95%CI: (0.8728, 0.9856))

2012 Annual Measure for Communities

| Hospitalizations per 1000 Summary | | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
|  | N | Mean | Std Dev | Median | Minimum | Maximum |
| Sample 1 | 327 | 287.46 | 55.03 | 286.67 | 151.40 | 468.17 |
| Sample 2 | 327 | 286.97 | 53.90 | 286.69 | 142.51 | 473.73 |

| Hospitalizations per 1000 Correlations | |
| --- | --- |
|  | Correlation (p-value) |
| Pearson Correlation | 0.99036 (<.0001) |
| Spearman Correlation (uses ranks) | 0.97720 (<.0001) |

|  | Hospitalizations per 1000 Quintile Comparison | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | Quintiles Sample 2 | | | | | |
|  |  | 1 | 2 | 3 | 4 | 5 | Total |
| Quintiles  Sample 1 | 1 | 61 | 5 | 0 | 0 | 0 | 66 |
| 2 | 5 | 48 | 12 | 0 | 0 | 65 |
| 3 | 0 | 12 | 46 | 8 | 0 | 66 |
| 4 | 0 | 0 | 8 | 50 | 7 | 65 |
| 5 | 0 | 0 | 0 | 7 | 58 | 65 |
|  | Total | 66 | 65 | 66 | 65 | 65 | 327 |

Quintile Agreement: 80% (263/327)

Weighted Kappa: 0.8777 (95%CI: (0.8491, 0.9063))

2012 Q1 Measure for States/Territories

| Hospitalizations per 1000 Summary | | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
|  | N | Mean | Std Dev | Median | Minimum | Maximum |
| Sample 1 | 53 | 73.02 | 12.87 | 75.81 | 37.69 | 92.41 |
| Sample 2 | 53 | 73.12 | 12.47 | 75.09 | 38.37 | 91.76 |

| Hospitalizations per 1000 Correlations | |
| --- | --- |
|  | Correlation (p-value) |
| Pearson Correlation | 0.99667 (<.0001) |
| Spearman Correlation (uses ranks) | 0.99540 (<.0001) |

|  | Hospitalizations per 1000 Quintile Comparison | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | Quintiles Sample 2 | | | | | |
|  |  | 1 | 2 | 3 | 4 | 5 | Total |
| Quintiles  Sample 1 | 1 | 11 | 0 | 0 | 0 | 0 | 11 |
| 2 | 0 | 10 | 1 | 0 | 0 | 11 |
| 3 | 0 | 1 | 8 | 1 | 0 | 10 |
| 4 | 0 | 0 | 1 | 10 | 0 | 11 |
| 5 | 0 | 0 | 0 | 0 | 10 | 10 |
|  | Total | 11 | 11 | 10 | 11 | 10 | 53 |

Quintile Agreement: 92% (49/53)

Weighted Kappa: 0.9528 (95%CI: (0.9067, 0.9989))

2012 Q1 Measure for Communities

| Hospitalizations per 1000 Summary | | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
|  | N | Mean | Std Dev | Median | Minimum | Maximum |
| Sample 1 | 327 | 76.70 | 14.86 | 77.48 | 37.55 | 122.73 |
| Sample 2 | 327 | 76.66 | 14.59 | 77.21 | 37.49 | 123.21 |

| Hospitalizations per 1000 Correlations | |
| --- | --- |
|  | Correlation (p-value) |
| Pearson Correlation | 0.95583 (<.0001) |
| Spearman Correlation (uses ranks) | 0.94905 (<.0001) |

|  | Hospitalizations per 1000 Quintile Comparison | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | Quintiles Sample 2 | | | | | |
|  |  | 1 | 2 | 3 | 4 | 5 | Total |
| Quintiles  Sample 1 | 1 | 60 | 6 | 0 | 0 | 0 | 66 |
| 2 | 6 | 41 | 16 | 1 | 1 | 65 |
| 3 | 0 | 17 | 39 | 10 | 0 | 66 |
| 4 | 0 | 1 | 11 | 46 | 7 | 65 |
| 5 | 0 | 0 | 0 | 8 | 57 | 65 |
|  | Total | 66 | 65 | 66 | 65 | 65 | 327 |

Quintile Agreement: 74% (243/327)

Weighted Kappa: 0.8318 (95%CI: (0.7967, 0.8669))

2012 Q2 Measure for States/Territories

| Hospitalizations per 1000 Summary | | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
|  | N | Mean | Std Dev | Median | Minimum | Maximum |
| Sample 1 | 53 | 68.01 | 11.54 | 69.73 | 36.88 | 84.69 |
| Sample 2 | 53 | 68.19 | 11.37 | 69.53 | 37.13 | 84.40 |

| Hospitalizations per 1000 Correlations | |
| --- | --- |
|  | Correlation (p-value) |
| Pearson Correlation | 0.99460 (<.0001) |
| Spearman Correlation (uses ranks) | 0.98976 (<.0001) |

|  | Hospitalizations per 1000 Quintile Comparison | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | Quintiles Sample 2 | | | | | |
|  |  | 1 | 2 | 3 | 4 | 5 | Total |
| Quintiles  Sample 1 | 1 | 11 | 0 | 0 | 0 | 0 | 11 |
| 2 | 0 | 9 | 2 | 0 | 0 | 11 |
| 3 | 0 | 2 | 7 | 1 | 0 | 10 |
| 4 | 0 | 0 | 1 | 8 | 2 | 11 |
| 5 | 0 | 0 | 0 | 2 | 8 | 10 |
|  | Total | 11 | 11 | 10 | 11 | 10 | 53 |

Quintile Agreement: 81% (43/53)

Weighted Kappa: 0.8820 (95%CI: (0.8125, 0.9515))

2012 Q2 Measure for Communities

| Hospitalizations per 1000 Summary | | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
|  | N | Mean | Std Dev | Median | Minimum | Maximum |
| Sample 1 | 327 | 68.99 | 13.79 | 69.36 | 35.82 | 113.56 |
| Sample 2 | 327 | 68.68 | 13.91 | 68.20 | 35.68 | 118.46 |

| Hospitalizations per 1000 Correlations | |
| --- | --- |
|  | Correlation (p-value) |
| Pearson Correlation | 0.94215 (<.0001) |
| Spearman Correlation (uses ranks) | 0.93225 (<.0001) |

|  | Hospitalizations per 1000 Quintile Comparison | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | Quintiles Sample 2 | | | | | |
|  |  | 1 | 2 | 3 | 4 | 5 | Total |
| Quintiles  Sample 1 | 1 | 57 | 9 | 0 | 0 | 0 | 66 |
| 2 | 9 | 39 | 14 | 3 | 0 | 65 |
| 3 | 0 | 16 | 39 | 9 | 2 | 66 |
| 4 | 0 | 1 | 11 | 44 | 9 | 65 |
| 5 | 0 | 0 | 2 | 9 | 54 | 65 |
|  | Total | 66 | 65 | 66 | 65 | 65 | 327 |

Quintile Agreement: 71% (233/327)

Weighted Kappa: 0.8050 (95%CI: (0.7674, 0.8427))

2012 Q3 Measure for States/Territories

| Hospitalizations per 1000 Summary | | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
|  | N | Mean | Std Dev | Median | Minimum | Maximum |
| Sample 1 | 53 | 65.40 | 11.38 | 67.71 | 36.94 | 81.06 |
| Sample 2 | 53 | 65.39 | 11.47 | 66.28 | 35.81 | 81.37 |

| Hospitalizations per 1000 Correlations | |
| --- | --- |
|  | Correlation (p-value) |
| Pearson Correlation | 0.99568 (<.0001) |
| Spearman Correlation (uses ranks) | 0.99097 (<.0001) |

|  | Hospitalizations per 1000 Quintile Comparison | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | Quintiles Sample 2 | | | | | |
|  |  | 1 | 2 | 3 | 4 | 5 | Total |
| Quintiles  Sample 1 | 1 | 11 | 0 | 0 | 0 | 0 | 11 |
| 2 | 0 | 10 | 1 | 0 | 0 | 11 |
| 3 | 0 | 1 | 7 | 2 | 0 | 10 |
| 4 | 0 | 0 | 2 | 9 | 0 | 11 |
| 5 | 0 | 0 | 0 | 0 | 10 | 10 |
|  | Total | 11 | 11 | 10 | 11 | 10 | 53 |

Quintile Agreement: 89% (47/53)

Weighted Kappa: 0.9292 (95%CI: (0.8729, 0.9855))

2012 Q3 Measure for Communities

| Hospitalizations per 1000 Summary | | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
|  | N | Mean | Std Dev | Median | Minimum | Maximum |
| Sample 1 | 327 | 68.99 | 13.79 | 69.36 | 35.82 | 113.56 |
| Sample 2 | 327 | 68.68 | 13.91 | 68.20 | 35.67 | 118.46 |

| Hospitalizations per 1000 Correlations | |
| --- | --- |
|  | Correlation (p-value) |
| Pearson Correlation | 0.95470 (<.0001) |
| Spearman Correlation (uses ranks) | 0.95161 (<.0001) |

|  | Hospitalizations per 1000 Quintile Comparison | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | Quintiles Sample 2 | | | | | |
|  |  | 1 | 2 | 3 | 4 | 5 | Total |
| Quintiles  Sample 1 | 1 | 59 | 7 | 0 | 0 | 0 | 66 |
| 2 | 7 | 43 | 14 | 1 | 0 | 65 |
| 3 | 0 | 11 | 42 | 13 | 0 | 66 |
| 4 | 0 | 4 | 9 | 42 | 10 | 65 |
| 5 | 0 | 0 | 1 | 9 | 55 | 65 |
|  | Total | 66 | 65 | 66 | 65 | 65 | 327 |

Quintile Agreement: 74% (241/327)

Weighted Kappa: 0.8242 (95%CI: (0.7884, 0.8599))

2012 Q4 Measure for States/Territories

| Hospitalizations per 1000 Summary | | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
|  | N | Mean | Std Dev | Median | Minimum | Maximum |
| Sample 1 | 53 | 66.51 | 11.45 | 68.96 | 35.95 | 81.83 |
| Sample 2 | 53 | 66.60 | 10.97 | 68.06 | 35.89 | 82.24 |

| Hospitalizations per 1000 Correlations | |
| --- | --- |
|  | Correlation (p-value) |
| Pearson Correlation | 0.99448 (<.0001) |
| Spearman Correlation (uses ranks) | 0.99234 (<.0001) |

|  | Hospitalizations per 1000 Quintile Comparison | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | Quintiles Sample 2 | | | | | |
|  |  | 1 | 2 | 3 | 4 | 5 | Total |
| Quintiles  Sample 1 | 1 | 10 | 1 | 0 | 0 | 0 | 11 |
| 2 | 1 | 9 | 1 | 0 | 0 | 11 |
| 3 | 0 | 1 | 8 | 1 | 0 | 10 |
| 4 | 0 | 0 | 1 | 9 | 1 | 11 |
| 5 | 0 | 0 | 0 | 1 | 9 | 10 |
|  | Total | 11 | 11 | 10 | 11 | 10 | 53 |

Quintile Agreement: 85% (45/53)

Weighted Kappa: 0.9056 (95%CI: (0.8432, 0.9680))

2012 Q4 Measure for Communities

| Hospitalizations per 1000 Summary | | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
|  | N | Mean | Std Dev | Median | Minimum | Maximum |
| Sample 1 | 327 | 70.38 | 13.87 | 70.35 | 33.26 | 129.33 |
| Sample 2 | 327 | 70.03 | 13.24 | 70.09 | 35.46 | 113.00 |

| Hospitalizations per 1000 Correlations | |
| --- | --- |
|  | Correlation (p-value) |
| Pearson Correlation | 0.94469 (<.0001) |
| Spearman Correlation (uses ranks) | 0.94571 (<.0001) |

|  | Hospitalizations per 1000 Quintile Comparison | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | Quintiles Sample 2 | | | | | |
|  |  | 1 | 2 | 3 | 4 | 5 | Total |
| Quintiles  Sample 1 | 1 | 58 | 7 | 1 | 0 | 0 | 66 |
| 2 | 8 | 41 | 14 | 2 | 0 | 65 |
| 3 | 0 | 14 | 35 | 17 | 0 | 66 |
| 4 | 0 | 3 | 16 | 38 | 8 | 65 |
| 5 | 0 | 0 | 0 | 8 | 57 | 65 |
|  | Total | 66 | 65 | 66 | 65 | 65 | 327 |

Quintile Agreement: 70% (229/327)

Weighted Kappa: 0.8012 (95%CI: (0.7634, 0.8390))

**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?*)

Kappa Statistic Interpretation

The Kappa statistic is a statistical measure of inter-rater reliability. It generally ranges from 0.0 to 1.0, where values near one are indicative of high reliability and values near zero indicate that agreement can be attributed to chance.

Landis and Koch, 1977 offer the following classification of Kappa interpretation:

|  |  |
| --- | --- |
| Kappa Range | Interpretation |
| <0 | Poor agreement |
| 0.00-0.20 | Slight agreement |
| 0.21-0.40 | Fair agreement |
| 0.41-0.60 | Moderate agreement |
| 0.61-0.80 | Substantial agreement |
| 0.81-1.00 | Almost perfect agreement |

The high correlation coefficients along with the high kappa statistics for quintile agreement, suggest high reliability for the annual and quarterly Hospitalizations per 1000 Beneficiaries measures when computed at the state/territory level.

The correlation coefficients along with the high kappa for quintile agreement, suggest high reliability for the annual and quarterly Hospitalizations per 1000 Beneficiaries measures when computed at the community level.

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

**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)*

The Commonwealth Fund report “Commission on a High Performance Health System, March 2012” contains results on a variety of measures, including Medicare Hospitalizations for Ambulatory Care-Sensitive Conditions per 100,000 beneficiaries. This measure is one of four measures in the Potentially Avoidable Hospital Use and Cost dimension. The report computes metrics for all 306 Hospital Referral Regions (HRRs) in the US and displays quartile (1, 2, 3, and 4) classification for each HRR for this dimension. Potentially Avoidable Hospital Use and Cost dimension quartile should correlate with Hospitalizations per 1,000 Medicare Beneficiaries measure.

Annual Measure: We computed Hospitalizations per 1,000 Beneficiaries for each HRR in the US and split the HRRs into quartiles. Quartiles for Commonwealth’s Use/Cost dimension and Hospitalizations per 1,000 Beneficiaries were compared.

Quarterly Measure: Since the annual measure was found to be valid, we compared the quarterly measure rates to the annual measure rate to test their validity.  In addition, the sum of the four quarterly measures for a year should be nearly identical to the annual measure for the same set of quarters.

Face validity: Gawande (2009, New Yorker) discussed the high costs of care in McAllen, TX, indicating that only Miami, FL had higher costs.  The article also explored reasons why El Paso, TX, just up the coast from McAllen, TX had quite a bit lower costs.  Finally, the article discussed the Grand Junction, CO area due to its very low costs.  We identified ZIP Codes for these four communities and computed our 2012 annual measure for them.

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

|  | Commonwealth Quartile vs. Hospitalization per 1000 Quartile | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
|  |  | Hospitalizations per 1000  Quartile | | | | |
|  | Frequency | 1 | 2 | 3 | 4 | Total |
| Commonwealth Quartile | 1 | 57 | 19 | 2 | 0 | 78 |
| 2 | 18 | 37 | 18 | 5 | 78 |
| 3 | 2 | 18 | 33 | 21 | 74 |
| 4 | 0 | 2 | 24 | 50 | 76 |
| Total | 77 | 76 | 77 | 76 | 306 |

HRR Commonwealth Potentially Avoidable Hospital Use and Cost Quartile and Hospitalizations per 1000 Beneficiaries quartile agreement is 177/306 = 58%.

Weighted Kappa statistic for quartile agreement is 0.63 (95%CI: (0.58, 0.69)).

The correlation of the 53 state/territory quarterly rates for 2012 Q1, 2012 Q2, 2012 Q3, and 2012 Q4 with the annual rate for 2012 were 0.996, 0.998, 0.997, and 0.997, respectively.

The average of the 53 state/territory rates are:

2012 Annual: 273.57

2012 Q1: 73.12

2012 Q2: 68.15

2012 Q3: 65.56

2012 Q4: 66.84

Sum of four 2012 quarters: 273.67

The correlation of the 327 community quarterly rates for 2012 Q1, 2012 Q2, 2012 Q3, and 2012 Q4 with the annual rate for 2012 were 0.986, 0.987, 0.991, and 0.981, respectively.

The average of the 327 rates are:

2012 Annual: 287.45

2012 Q1:76.67

2012 Q2: 71.56

2012 Q3: 68.89

2012 Q4: 70.43

Sum of four 2012 quarters: 287.55

Face Validity:

| Community | Hospitalizations per 1000 Benes |
| --- | --- |
| McAllen | 318.54 |
| Miami | 378.30 |
| El Paso | 274.41 |
| Grand Junction | 189.25 |

In summary, the average of the 53 state/territory 2012 hospitalizations per 1000 beneficiaries for blacks was 329.47, compared to 266.82 for whites.  Also, for age groups 65-69, 70-74, 75-79, 80-84< 85+ hospitalizations per 1000 were 162.22, 217.60, 285.43, 359.62, and 456.36, respectively.

**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 percent agreement and kappa statistic support the validity of the Hospitalizations per 1,000 Beneficiaries measure.

The quarterly rates have a very high correlation with the annual rate and the sum of the four quarterly rates is nearly identical to the annual rate.  Therefore, since testing supports validity of the annual measure and the quarterly measures correlate well and sum to the annual measure, the quarterly measures also appear to be valid.

Face Validity:

Based on New Yorker article discussing high cost, we would expect the same order for the four communities, namely Miami , McAllen, El Paso, and Grand Junction.  Our 2012 measure rate found Miami to be the highest, McAllen second, El Paso, third, and Grand Junction a distant fourth, matching exactly what was reported for cost in the Anew Yorker article.

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

**2b3. EXCLUSIONS ANALYSIS**

**NA** ☒ **no exclusions — *skip to section*** [***2b4***](#section2b4)

**2b3.1. Describe the method of testing exclusions and what it tests** (*describe the steps―do not just name a method; what was tested, e.g., whether exclusions affect overall performance scores; what statistical analysis was used*)

**2b3.2. What were the statistical results from testing exclusions**? (*include overall number and percentage of individuals excluded, frequency distribution of exclusions across measured entities, and impact on performance measure scores*)

**2b3.3. What is your interpretation of the results in terms of demonstrating that exclusions are needed to prevent unfair distortion of performance results?** (*i.e., the value outweighs the burden of increased data collection and analysis.*  *Note:* ***If patient preference is an exclusion****, the measure must be specified so that the effect on the performance score is transparent, e.g., scores with and without exclusion*)

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

**2b4. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES**  
***If not an intermediate or health outcome, or PRO-PM, or resource use measure, skip to section*** [***2b5***](#section2b5)***.***

**2b4.1. What method of controlling for differences in case mix is used?**

☐ **No risk adjustment or stratification**

☐ **Statistical risk model with** Click here to enter number of factors **risk factors**

☐ **Stratification by** Click here to enter number of categories **risk categories**

☒ **Other,** Seasonal Adjustment for the quarterly measure

**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**.

There is no risk adjustment for differences in case mix. The purpose of this measure is to support population health improvement initiatives by tracking change in hospitalization rates for a population over time, and not to compare hospitalization rates among different populations at a single point in time.   The population of interest is all FFS Medicare beneficiaries residing in the selected geographic subdivision (states, zip code, designations of community, etc), which is unlikely to change with regard to important characteristics over short intervals removing the need for risk adjustment. Although other measures of medical utilization are typically adjusted by patient characteristics, this measure is designed to be useful to communities to monitor progress in hospitalization reduction, and important characteristics of communities that are associated with the capacity to change medical practice. Because there is no defined set of variables that characterize a community’s capacity to change its medical utilization practices, improvement science is grounded in the assumption that complex systems such as communities are dynamic before, during and after intervention tests, and that any given system characteristic may or may not endure throughout the intervention period. Therefore, we do not adjust for any patient characteristics for either the annual or quarterly measures.

While we do not risk adjust for patient characteristics we do incorporate a seasonal adjustment when trending the quarterly measure. This allows for comparison of any and all quarters (e.g., Q1 2011; Q2 2011; Q3 2012) and trending for any state/territory or community. Without the adjustment only like quarters (e.g., Q1 2010 and Q1 2011) can be compared. Please refer to S.14 and S2.b for details.

**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*)

N/A

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

**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*)

N/A

*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*)**:**

N/A

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

N/A

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

N/A

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

N/A

**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*)

N/A

**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*)

N/A

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

**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)*

Annual Measure: Relative improvement was computed comparing 2010 to 2012 using the annual rate for states/territories and communities.

Quarterly Measure: Control charts were used to assess improvement and deterioration for state/territory and community seasonally adjusted quarterly rates. Quarter 1, 2009 through quarter 3, 2011 were used to compute the center line and control limits. “Out-of-control” points were assessed using Quarter 4, 2011 through Quarter 1, 2013.

“Out-of-control” Rules

1. Any point during the intervention period that is above the upper control limit (deterioration) or below the lower control limit (improvement).
2. Eight consecutive points above the center line (deterioration) or eight consecutive points below the center line (improvement).
3. If states/territories and communities are flagged for both deterioration and improvement we identify them based on whichever (deterioration or improvement) occurred later.

**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*)

Quarterly: Of the 53 states/territories 51 (96%) demonstrated improvement and 0 (0%) demonstrated deterioration. Of the 327 communities, 263 (80%) demonstrated improvement and 4 (1%) demonstrated deterioration.

States/Territories:

| Relative Improvement in Hospitalizations per 1000  (2010 annual to 2012 annual) | | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
| N | Minimum | Maximum | Mean | 25th Pctl | 50th Pctl | 75th Pctl |
| 53 | -0.50 | 11.81 | 6.54 | 5.54 | 6.93 | 8.07 |

Communities:

| Relative Improvement in Hospitalizations per 1000  (2010 annual to 2012 annual) | | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
| N | Minimum | Maximum | Mean | 25th Pctl | 50th Pctl | 75th Pctl |
| 327 | -15.93 | 20.73 | 6.52 | 3.56 | 6.55 | 9.39 |

**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?*)

The results suggest the measure is sensitive to the effects of interventions intended to reduce hospitalizations.

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

**2b6. COMPARABILITY OF PERFORMANCE SCORES WHEN MORE THAN ONE SET OF SPECIFICATIONS**

***If only one set of specifications, this section can be skipped.***

**Note***: This criterion is directed to measures with more than one set of specifications/instructions (e.g., one set of specifications for how to identify and compute the measure from medical record abstraction and a different set of specifications for claims or eMeasures). It does not apply to measures that use more than one source of data in one set of specifications/instructions (e.g., claims data to identify the denominator and medical record abstraction for the numerator).* ***If comparability is not demonstrated, the different specifications should be submitted as separate measures.***

**2b6.1. Describe the method of testing conducted to demonstrate comparability of performance scores for the same entities across the different data sources/specifications** (*describe the steps―do not just name a method; what statistical analysis was used*)

N/A

**2b6.2. What were the statistical results from testing comparability of performance scores for the same entities when using different data sources/specifications?** (*e.g., correlation, rank order*)

N/A

**2b6.3. What is your interpretation of the results in terms of demonstrating comparability of performance measure scores for the same entities across the different data sources/specifications?** (i*.e., what do the results mean and what are the norms for the test conducted*)

N/A

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

**2b7. MISSING DATA ANALYSIS AND MINIMIZING BIAS**

**2b7.1. Describe the method of testing conducted to identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased** due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias (*describe the steps―do not just name a method; what statistical analysis was used*)

N/A

**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*)

N/A

**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*)

N/A