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

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

**Measure Title**: Hospital 30-Day All-Cause Risk-Standardized Readmission Rate (RSRR) following Vascular Procedures

**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*).

The datasets used for testing included Medicare Part A inpatient and outpatient claims, Part B claims, as well as the Medicare Enrollment Database (EDB). The dataset used varies by testing type; see Section 1.7 for details.

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

January 1, 2009-December 31, 2010. The dates used vary by testing type; see Section 1.7 for details.

**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 this measure, hospitals are the measured entities. All non-federal, acute inpatient US hospitals (including territories) with Medicare FFS beneficiaries over the age of 65 are included. The number of measured entities (hospitals) varies by testing type; see Section 1.7 for details.

**1.6. How many and which patients were included in the testing and analysis (by level of analysis and data source)**? (*identify the number and descriptive characteristics of patients included in the analysis (e.g., age, sex, race, diagnosis); if a sample was used, describe how patients were selected for inclusion in the sample*)

The number of patients varies by testing type; see Section 1.7 for details

**1.7. If there are differences in the data or sample used for different aspects of testing (e.g., reliability, validity, exclusions, risk adjustment), identify how the data or sample are different for each aspect of testing reported below**.

The datasets used for testing included Medicare Part A inpatient and outpatient claims, Part B claims, as well as the Medicare Enrollment Database (EDB). The datasets, dates, number of measured entities and number of hospital stays used in each type of testing are as follows:

Dataset A:  
Dates: January 1 – December 31, 2009 (100%)

Number of hospitals: 2,662

Number of hospital stays: 253,956

Descriptive characteristics of patients: Average age of 76.5 years, 54.7% male

Dataset used for: Development (cohort and outcome definition); Exclusions analysis; Risk adjustment; Identification of statistically significant & meaningful differences in performance

Dataset A1:

Dates: January 1 – December 31, 2009 (50%)

Number of hospitals: 2,479

Number of hospital stays: 126,971

Descriptive characteristics of patients: Average age of 76.5 years, 54.7% male

Dataset used for: Development (variable selection); Reliability testing; Risk adjustment

Dataset A2:

Dates: January 1 – December 31, 2009 (50%)

Number of hospitals: 2,502

Number of hospital stays: 126,985

Descriptive characteristics of patients: Average age of 76.5 years, 54.6% male

Dataset used for: Reliability testing; Risk adjustment

Dataset R:

Dates: January 1 – December 31, 2010 (100%)

Number of hospitals: 2,623

Number of hospital stays: 248,337

Descriptive characteristics of patients: Average age of 76.6 years, 54.9% male

Dataset used for: Reliability testing; Risk adjustment; Identification of statistically significant & meaningful differences in performance

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

**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*)

Data element reliability

In constructing measures we aim to utilize only those data elements from the claims that have both face validity and reliability. We avoid the use of fields that are thought to be coded inconsistently across hospitals or providers. Specifically, we use fields that are consequential for payment and which are audited. We identify such variables through empiric analyses and our understanding of CMS auditing and billing policies, and seek to avoid variables which do not meet this standard. For example, “discharge disposition” is a variable in Medicare claims data that is not thought to be a reliable variable for identifying a transfer between two acute care facilities. Thus, we derive a variable using admission and discharge dates as a surrogate for “discharge disposition” to identify hospital stays involving transfers. This allows us to identify these stays using variables in the claims data which have greater reliability than the “discharge disposition” variable.

In addition, CMS has in place several hospital auditing programs used to assess overall claims code accuracy, to ensure appropriate billing, and for overpayment recoupment. CMS routinely conducts data analyses to identify potential problem areas and detect fraud, and audits important data fields used in our measures, including diagnosis and procedure codes and other elements that are consequential to payment.

Finally, we assess the reliability of the data elements by comparing risk factor frequencies and odds ratios (ORs) and corresponding confidence intervals (CIs) from logistic regression models in three datasets (Datasets A1, A2, and R).

Measure score reliability

The reliability of a measurement is the degree to which repeated measurements of the same entity agree with each other. For measures of hospital performance, the measured entity is naturally the hospital, and reliability is the extent to which repeated measurements of the same hospital give similar results. In line with this thinking, our approach to assessing reliability is to consider the extent to which assessments of a hospital using different but randomly selected subsets of patients produce similar measures of hospital performance. That is, we take a “test-retest” approach in which hospital performance is measured once using a random subset of patients, and then measured again using the remaining subset of patients, and the agreement of the two resulting performance measures is compared across hospitals (Rousson et al. 2002).

For test-retest reliability, we combined index hospital stays from successive measurement periods into one dataset, randomly sampled half of patients within each hospital, calculated the measure for each hospital, and repeated the calculation using the second half. Thus, each hospital is measured twice, but each measurement is made using an entirely distinct set of patients. To the extent that the calculated measures of these two subsets agree, we have evidence that the measure is assessing an attribute of the hospital, not of the patients. As a metric of agreement, we calculated the intra-class correlation coefficient (Shrout et al. 1979) and assessed the values according to conventional standards (Landis et al. 1977).Specifically, we used a combined 2009-2010 sample, randomly split it into two approximately equal subsets of patients, and calculated the RSRR for each hospital for each sample. The agreement of the two RSRRs was quantified using the intra-class correlation as defined by ICC(2,1) by Shrout and Fleiss (1979).

Using two independent samples provides a stringent estimate of the measure’s reliability, compared with using two random but potentially overlapping samples which would exaggerate the agreement. Moreover, because our final measure is derived using hierarchical logistic regression, and a known property of hierarchical logistic regression models is that smaller volume hospitals contribute less ‘signal’, a split sample using a single measurement period would introduce extra noise, potentially underestimating the actual test-retest reliability that would be achieved if the measure were reported using two years of data.

Rousson V, Gasser T, Seifert B. "Assessing intrarater, interrater and test–retest reliability of continuous measurements," *Statistics in Medicine,* 2002, 21:3431-3446.

Shrout P, Fleiss J. Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin*, 1979, 86, 420-3428.

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

**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*)

Data element reliability results:

No notable differences were observed in risk factor frequency or odds ratios across the three datasets.

Measure score reliability results:

There were 502,293 admissions in the combined two-year sample, with 251,519 and 250,774 in the two randomly selected samples, respectively. The ICC score, indicating agreement between the two independent assessments of each hospital, was 0.40.

**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 stability over time of the risk factor frequencies and odds ratios suggests that the underlying data elements are reliable. Additionally, the ICC score demonstrates moderate agreement across samples using a “strict” approach to assessment and would likely improve with greater sample size.

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

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

**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)*

Measure validity is demonstrated through prior validity testing done on our claims-based measures, through use of established measure development guidelines, and by systematic assessment of measure face validity by a TEP of national experts and stakeholder organizations.

Validity of Claims-Based Measures:

Our team has demonstrated for a number of prior measures the validity of claims-based measures for profiling hospitals by comparing either the measure results or individual data elements against medical records. CMS validated six NQF-endorsed measures currently in public reporting (AMI, heart failure, and pneumonia mortality and readmission) with models that used chart-abstracted data for risk-adjustment. Specifically, claims model validation was conducted by building comparable models using abstracted medical chart data for risk-adjustment for heart failure patients (National Heart Failure data) (Krumholz et al. 2006; Keenan et al. 2008) AMI patients (Cooperative Cardiovascular Project data) (Krumholz, Wang, et al. 2006) and pneumonia patients (National Pneumonia Project dataset) (Bratzler et al. 2011). When both models were applied to the same patient population, the hospital risk-standardized rates estimated using the claims-based risk adjustment models had a high level of agreement with the results based on the medical record model, thus supporting the use of the claims-based models for public reporting.

We have also completed two national, multi-site validation efforts for two procedure-based complications measures (for primary elective hip/knee arthroplasty and implantable cardioverter defibrillator (ICD)). Both projects demonstrated strong agreement between complications coded in claims and abstracted medical chart data.

Validity Indicated by Established Measure Development Guidelines:

We developed this measure in consultation with national guidelines for publicly reported outcomes measures, with outside experts, and with the public. The measure is consistent with the technical approach to outcomes measurement set forth in NQF guidance for outcomes measures (National Quality Forum, 2010), CMS Measure Management System (MMS) guidance, and the guidance articulated in the American Heart Association scientific statement, “Standards for Statistical Models Used for Public Reporting of Health Outcomes” (Krumholz, Brindis, et al. 2006).

Validity as Assessed by External Groups:

Throughout measure development, we obtained expert and stakeholder input via three mechanisms: regular discussions with an advisory working group, a national Technical Expert Panel (TEP), and a 30-day public comment period in order to increase transparency and to gain broader input into the measure.

The working group was assembled, and regular meetings were held throughout the development phase. The working group was tailored for development of this measure and consisted of clinicians (including a noninvasive cardiologist, vascular surgeon, family physician, and interventional cardiologist), and other professionals with expertise in biostatistics, measure methodology, and quality improvement. Working group meetings addressed key issues related to measure development, including weighing the pros and cons of and finalizing key decisions (e.g., defining the measure cohort and outcome) to ensure the measure is meaningful, useful, and well-designed. The working group provided a forum for focused expert review and discussion of technical issues during measure development prior to consideration by the broader TEP.

In addition to the working group, and in alignment with the CMS MMS, we convened a TEP to provide input and feedback during measure development from a group of recognized experts in relevant fields. To convene the TEP, we released a public call for nominations and selected individuals to represent a range of perspectives, including physicians, consumers, and purchasers, as well as individuals with experience in quality improvement, performance measurement, and health care disparities. We held three structured TEP conference calls consisting of presentation of key issues, our proposed approach, and relevant data, followed by open discussion among TEP members.

Following completion of the preliminary model, we solicited public comment on the measure through the CMS site link <https://www.CMS.gov/MMS/17_CallforPublicComment.asp>. The public comments were then posted publicly for 30 days. The resulting input was taken into consideration during the final stages of measure development and contributed to minor modifications to the measure.

Face Validity as Determined by TEP:

One means of confirming the validity of this measure was face validity assessed by our Technical Expert Panel (TEP), which included 12 members including vascular surgeons, interventional cardiologists, interventional radiologists, and other experts in quality assessment.

List of TEP Members

1)Terry Golash, MD; Aetna, Senior Medical Director; New York, NY

2) Bruce Hall, MD, PhD, MBA; Washington University Saint Louis, Professor of Surgery and Healthcare Management; St. Louis, MO

3) Jeffrey Indes, MD; Yale University School of Medicine, Assistant Professor of Surgery and Radiology, Section of Vascular Surgery; New Haven, CT

4) Sanjay Misra, MD; Mayo Clinic, Associate Professor of Radiology; Rochester, MN

5) Leila Mureebe, MD; Duke University Medical Center, Assistant Professor of Surgery; Attending Surgeon; Durham, NC

6) Ileana L. Piña, MD, MPH; Case Western Reserve University, Professor of Medicine & Professor of Epidemiology/Biostatistics; Cleveland Heights, OH

7) Anne Roberts, MD; University of California, San Diego, Professor of Radiology; Chief of Vascular and Interventional Radiology; La Jolla, CA

8) Sean P. Roddy, MD; Society for Vascular Surgery, Health Policy Committee Chair; Albany, NY

9) John Santa, MD, MPH; Consumer Reports Health Ratings Center, Director; Yonkers, NY

10) Laurel Trujillo, MD; Palo Alto Medical Foundation, Medical Director of Quality; Mountain View, CA

11) Todd R. Vogel, MD; UMDNJ-RWJ Medical School, Assistant Professor of Surgery, Director of Vascular Laboratory RWJUH, Director of Surgical Outcomes Research Group; New Brunswick, NJ

12) Christopher J. White, MD; Ochsner Clinic Foundation Department of Cardiology, Chairman; New Orleans, LA

We systematically assessed the face validity of the measure score as an indicator of quality by soliciting the TEP members’ agreement with the following statement: “*The risk-standardized readmission rates obtained from the vascular readmission measure as specified will provide an accurate reflection of quality.”*

The 11 attending TEP members agreed with the face validity of the measure, with only one expressing any concern with a vote of 3 on a six-point scale: 1=Strongly disagree, 2=Moderately disagree, 3=Somewhat disagree, 4=Somewhat agree, 5=Moderately agree, 6=Strongly agree.

**ICD-9 to ICD-10 Conversion**

Statement of Intent and Process of Conversion:

This application includes ICD-10 codes that correspond to the ICD-9 codes included in our measure specifications. The goal of conversion to ICD-10 was to convert this measure to a new code set, fully consistent with the intent of the original measure. ICD-10 codes were initially identified using GEM mapping software. We then enlisted the help of clinicians with expertise in relevant areas to select and evaluate which ICD-10 codes map to the ICD-9 codes currently in use for this measure.  An ICD-9 to ICD-10 crosswalk is attached in field S.2b (Data Dictionary).

Citations:

Krumholz HM, Wang Y, Mattera JA, et al. An administrative claims model suitable for profiling hospital performance based on 30-day mortality rates among patients with an acute myocardial infarction. *Circulation* 2006;113(13):1683-92.

Krumholz HM, Wang Y, Mattera JA, et al. An administrative claims model suitable for profiling hospital performance based on 30-day mortality rates among patients with heart failure. *Circulation* 2006;113:1693-1701.

Bratzler DW, Normand SL, Wang Y, et al. An administrative claims model for profiling hospital 30-day mortality rates for pneumonia patients. *PLoS One* 2011;6(4):e17401.

Keenan PS, Normand SL, Lin Z, et al. An administrative claims measure suitable for profiling hospital performance on the basis of 30-day all-cause readmission rates among patients with heart failure. *Circulation* 2008;1(1):29-37.

National Quality Forum. National voluntary consensus standards for patient outcomes, first report for phases 1 and 2: A consensus report <http://www.qualityforum.org/projects/Patient_Outcome_Measures_Phases1-2.aspx>. Accessed August 19, 2010.

Krumholz HM, Brindis RG,Brush JE, et al. Standards for Statistical Models Used for Public Reporting of Health Outcomes: An American Heart Association Scientific Statement From the Quality of Care and Outcomes Research Interdisciplinary Writing Group: Cosponsored by the Council on Epidemiology and Prevention and the Stroke Council Endorsed by the American College of Cardiology Foundation. *Circulation.* January 24, 2006 2006;113(3):456-462.

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

The results of the TEP rating of agreement with the validity statement were as follows:

N=11

Mean rating=4.6

**Frequency of Ratings of Agreement**

|  |  |
| --- | --- |
| **# of Responses** | **Rating** |
| 0 | 1 (Strongly disagree) |
| 0 | 2 (Moderately disagree) |
| 1 | 3 (Somewhat disagree) |
| 3 | 4 (Somewhat agree) |
| 6 | 5 (Moderately agree) |
| 1 | 6 (Strongly agree) |

**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?*)

These face validity testing results demonstrate TEP agreement with overall face validity of the measure as specified.

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

**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*)

Exclusions were those determined by expert input to be clinically relevant. These exclusions are consistent with similar NQF-endorsed readmission measures. Rationales for the exclusions are detailed in Denominator Exclusions section (S.9). To ascertain impact of exclusions on the cohort, we examined overall frequencies and proportions of the total cohort excluded for exclusions that are not data requirements (such that without the data, measure calculation would not be possible), or have minimal impact on the measure due to very low frequency.

**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*)

For the purposes of tabulation, exclusions are performed sequentially. Thus, a hospital stay that would be excluded based on multiple criteria is counted in the first criterion only.

Among 1,691 hospitals with at least 25 index stays in 2009:

|  |  |  |  |
| --- | --- | --- | --- |
| **Exclusion** | **N** | **%** | **Distribution across hospitals: Min, 25th, 50th, 75th percentile, max** |
| 1. Hospital stays for patients without at least 30 days post‐discharge information | 745 | 0.3% | n/a (data-related exclusion) |
| 2. Hospital stays in which patients leave hospital against medical advice (AMA) | 225 | 0.08% | n/a (low impact) |
| 3. Hospital stays with a qualifying vascular procedure that occur within 30 days of a previous hospital stay with a qualifying vascular procedure | 12,444 | 4.6% | (0, 2.6, 4.0, 6.1, 27.6) |

These exclusions represent 5.0% of the initial cohort (n= 267,370).

**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 listed above were based on clinical input or are required for the determination of the outcome. Exclusions 1 and 3 are needed for valid calculation of the measure. Exclusion 2 (patients who leave AMA) is needed for acceptability of the measure to hospitals, who do not have the opportunity to adequately prepare such patients for discharge. Because of a very small percent of patients being excluded it is unlikely to affect measure score.

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

**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** 46 **risk factors**

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

**Other,** Click here to enter description

**2b4.2. If an outcome or resource use measure is not risk adjusted or stratified, provide rationale and analyses to demonstrate that controlling for differences in patient characteristics (case mix) is not needed to achieve fair comparisons across measured entities**.

N/A. 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*)

Our goal was to develop a parsimonious model that includes clinically relevant variables strongly associated with risk of readmission. The candidate variables for the model were derived from the index hospital stay, with comorbidities identified from the index hospital stay secondary diagnoses (excluding potential complications), 12 months pre-index hospital stay inpatient Part A data, outpatient hospital data, and Part B physician visit data.

For candidate variable selection using Dataset A1 (see item 1.7), we started with the 189 diagnostic groups included in the Hierarchical Condition Category (HCC) clinical classification system [1]. The HCC clinical classification system was developed for CMS in preparation for all-encounter risk adjustment for Medicare Advantage (managed care) plans and represented a refinement of an earlier risk-adjustment method based solely on principal inpatient diagnosis. The HCC model makes use of all physician and hospital encounter diagnoses and was designed to predict a beneficiary’s expenditures based on the total clinical profile represented by all of his/her assigned HCCs. Under the HCC algorithm, the 15,000+ ICD-9-CM diagnosis codes are first assigned to one of 804 mutually exclusive groupings (“DxGroups”) and then subsequently aggregated into 189 condition categories (CCs). We used the ICD-9-to-CC assignment map, which is maintained by CMS. To select candidate variables, a team of clinicians reviewed all 189 CCs and excluded those that were not relevant to the Medicare population or that were not clinically relevant to the readmission outcome (e.g., attention deficit disorder, female infertility). Clinically relevant CCs were selected as candidate variables; some of these CCs were combined into clinically coherent groups. For each CC, the team determined whether the particular condition might represent a complication of care that developed during the hospital stay and was not present at the time of arrival at the hospital. A list of CCs that were considered as potential complications, and were thus not included in the risk adjustment if coded *only* during the index hospital stay, is presented in the attached appendix (measure methodology report, Appendix G). Other candidate variables included age, gender, and procedure group (e.g., limb-open, head/neck-endovascular). All candidate variables are listed in item 2b4.4.

To inform variable selection a modified approach to stepwise logistic regression was performed. Dataset A1 was used to create 500 bootstrap samples. For each sample, we ran a logistic regression that included all candidate variables. The results were summarized to show the percentage of times that each of the candidate variables was significantly associated with readmission (at the p<0.05 level) in the 500 bootstrap samples (e.g., 70% would mean that the candidate variable was selected as significant at p<0.05 in 70% of the estimations). We also assessed the direction and magnitude of the regression coefficients.

The working group reviewed these results and decided to retain all risk adjustment variables above a 70% cutoff. The 70% cutoff was chosen because variables above this cutoff demonstrate a relatively strong association with readmission, and were clinically relevant. The one exception was cataract, which despite a high frequency of being associated with readmission in the 500 bootstrap samples (99.8%) was not considered clinically relevant. Variables which were significantly associated with readmission in less than 70% of the bootstrap samples were only included in the final model if they were markers for end of life/frailty or they were on the same clinical spectrum as a variable above the 70% cutoff and were clinically important for vascular procedure patients. This resulted in a final risk-adjusted readmission model that included 46 variables (see item 2b4.4).

[1] Pope, G., et al., Principal Inpatient Diagnostic Cost Group Models for Medicare Risk Adjustment. Health Care Financing Review, 2000. 21(3): p. 26.

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

Candidate and final model variables, with a corresponding CC to ICD-9 code map, are listed in the accompanying Excel file (tab 2b4.4).

**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*)

Model performance was assessed in three datasets (Datasets A1, A2, and R). For all three datasets, we computed three summary statistics for assessing model performance (Harrell 2001):

***Discrimination statistics:***

(1) Area under the receiver operating characteristic (ROC) curve (the c-statistic (also called ROC) is the probability that predicting the outcome is better than chance, which is a measure of how accurately a statistical model is able to distinguish between a patient with and without an outcome)

(2) Predictive ability (discrimination in predictive ability measures the ability to distinguish high-risk subjects from low-risk subjects. Therefore, we would hope to see a wide range between the lowest decile and highest decile)

***Calibration statistics:***

(3) Over-fitting indices (over-fitting refers to the phenomenon in which a model accurately describes the relationship between predictive variables and outcome in Dataset A1 used for development but fails to provide valid predictions in new patients)

F.E. Harrell and Y.C.T. Shih, Using full probability models to compute probabilities of actual interest to decision makers, *Int. J. Technol. Assess. Health Care* **17** (2001), pp. 17–26.

*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*)**:**

Dataset A1:

C statistic = 0.67

Predictive ability (lowest decile %, highest decile %): (6.4, 23.5)

Dataset A2:

C statistic = 0.67

Predictive ability (lowest decile %, highest decile %): (6.5, 23.3)

Dataset R:

C statistic = 0.67

Predictive ability (lowest decile %, highest decile %): (6.6, 23.8)

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

Dataset A1:

Calibration (over-fitting statistics): (0,1)

Dataset A2:

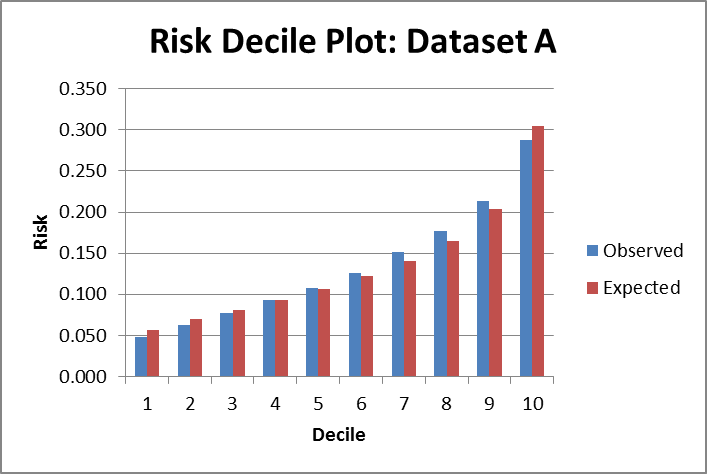
Calibration (over-fitting statistics): (-0.03, 0.98)

Dataset R:

Calibration (over-fitting statistics): (-0.03, 0.98)

**2b4.8. Statistical Risk Model Calibration – Risk decile (Hosmer-Lemeshow Test Decile) plots or calibration curves**:

The risk decile plot is a graphical depiction of the deciles calculated to measure predictive ability. Below, we present the risk decile plot for Dataset A, representative of risk decile plots for all other datasets, showing the distributions for the current measure cohort:



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

N/A. This measure is not risk stratified.

**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*)

***Discrimination Statistics***

The C-statistic of 0.67 was not substantially different across datasets and indicates good model discrimination.

The model indicated a wide range between the lowest decile and highest decile, indicating the ability to distinguish high-risk subjects from low-risk subjects.

***Calibration Statistics***

*Over-fitting (Calibration γ0, γ1)*

If the γ0 in the validation samples are substantially far from zero and the γ1 is substantially far from 1, there is potential evidence of over-fitting. The calibration value close to 0 at one end and close to 1 on the other end indicates good calibration of the model.

***Risk Decile Plots***

Higher deciles of the predicted outcomes are associated with higher observed outcomes, which shows a good calibration of the model. The risk decile plot indicates excellent discrimination of the model and good predictive ability.

***Overall Interpretation***

Interpreted together, our diagnostic results demonstrate the risk-adjustment model adequately controls for differences in patient characteristics (case mix).

**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)*

The method for discriminating hospital performance has not been determined. For publicly reported measures of hospital outcomes developed with similar methodology, CMS currently calculates an interval estimate for each risk-standardized rate to characterize the amount of uncertainty associated with the rate compares the interval estimate to the national crude rate for the outcome, and categorizes hospitals as “better than,” “worse than,” or “no different than” the US national rate. However, the decision to publicly report this measure and the approach to discriminating performance has not been determined.

See attached appendix (measure technical report, section 2.9) for description of analytic method.

**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*)

For data from 2009-2010 (Datasets A and R) with 2,870 providers, unadjusted hospital readmission rates range from 8.1%-16.7% (25th and 75th percentiles, respectively) with a median of 12.5%. The RSRRs range from 10.3%-18.9% with a median of 13.8%. The 25th and 75th percentile of the RSRRs is 13.2% and 14.0%, respectively.

**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?*)

These results suggest there are meaningful differences in the quality of care received for patients following hospitalization for a vascular procedure.

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

**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