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

*Last Updated 12/3/13*

**Measure Number** (*if previously endorsed*)**:** 0695

**Measure Title**: Hospital 30-Day Risk-Standardized Readmission Rates following Percutaneous Coronary Intervention (PCI)

**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: | other: |

**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 dataset used for testing included Medicare Part A claims, the National Cardiovascular Data Registry (NCDR) CathPCI Registry, and the Medicare Enrollment Database

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

The dates used vary by testing type; see Section 1.7 for details. The date range was 2006–2011.

**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) that participate in the American College of Cardiology (ACC) NCDR’s CathPCI Registry and care for Medicare Fee-for-Service (FFS) beneficiaries who are 65 years of age or older 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 admissions 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, dates, number of measured entities, and number of admissions used in each type of testing are as follows. For most testing requirements, we used data and analyses from the original submission to the National Quality Forum (NQF) of the PCI readmission measure. These analyses used a cohort of patients undergoing PCI in 2006-2007 for whom NCDR CathPCI Registry data had been successfully linked with corresponding administrative claims data. However, we also conducted additional analyses to meet newer testing requirements, and these analyses were performed using comparable linked data from 2010-2011. Details are provided below.

Reliability testing (Section 2a2) and exclusions testing (Section 2b3)

The measure reliability dataset linked the CathPCI and Medicare Part A claims data from 2010-2011. The combined two-year sample included 277,512 PCIs on Medicare FFS patients aged 65 years and older performed in 1,197 hospitals (mean age 75.15 years; % female=39.95%). We then randomly split the sample, leaving 138,756 admissions to 1,190 hospitals in one randomly selected sample and 138,756 admissions to 1,193 hospitals in the remaining sample for patients aged 65 years and older. After excluding hospitals with fewer than 25 cases in each sample, the first sample contained 970 hospitals and the second sample contained 969 hospitals. The linked dataset was also used for measure exclusions testing (Section 2b3).

For data element reliability, we utilized data and analyses from the original measure NQF submission as part of initial measure development. For these analyses, we identified PCI procedures in the CathPCI Registry in which the patient was released from the hospital between January and December 2007. This development sample consisted of 128,745 patient stays at 766 hospitals. For measure testing, we identified a cohort of PCIs in which the patient was released from the hospital between January and December 2006. This validation sample consisted of 117,375 patient stays at 618 hospitals.

Validity testing (Section 2b2)

Results of validity testing use data from the original measure NQF submission. For measure development, we identified PCI procedures in the CathPCI Registry in which the patient was released from the hospital between January and December 2007. This development sample consisted of 128,745 patient stays at 766 hospitals. For measure testing, we identified a cohort of PCIs in which the patient was released from the hospital between January and December 2006. This validation sample consisted of 117,375 patient stays at 618 hospitals.

Measure development and risk-adjustment dataset (Section 2b4)

In measure development, we identified PCI procedures in the CathPCI Registry in which the patient was released from the hospital between January and December 2007. We merged PCI admissions in the NCDR CathPCI Registry data and PCI admissions in Medicare claims data to derive cohorts for development using probabilistic matching methodology. There were 128,745 cases discharged from the 766 hospitals in the validation sample. This development sample had a crude readmission rate of 11.1%

For testing to identify meaningful differences in performance (Section 2b5)

The measure reliability dataset linked the CathPCI and Medicare Part A claims data from 2010-2011. The combined two-year sample included 277,512 to 1,197 hospitals with 138,756 admissions to 1,190 hospitals in one randomly selected sample and 138,756 admissions to 1,193 hospitals in the remaining sample for patients aged 65 years and older. After excluding hospitals with fewer than 25 cases in each sample, the first sample contained 970 hospitals and the second sample contained 969 hospitals. In addition to being used for reliability testing, the linked dataset was used for measure exclusions testing (Section 2b3).

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

**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 the measure 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 admissions involving transfers. This allows us to identify these admissions 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 analysis 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.

In addition, as an example of some of the methods that could be used to ensure data quality, we describe the NCDR’s existing Data Quality Program (DQP). The two main component of the DQP are complementary and consist of the Data Quality Report (DQR) and the Data Audit Program (DAP). The DQR process assesses the completeness and validity of the electronic data submitted by participating hospitals. Hospitals must achieve >95% completeness of specific data elements identified as ‘core fields’ to be included in the registry’s data warehouse for analysis. The ‘core fields’ include the variables included in 25 our risk adjustment models. The process is iterative, providing hospitals with the opportunity to correct errors and resubmit data for review and acceptance into the data warehouse. The DAP consists of annual on-site chart review and data abstraction. Among participating hospitals that pass the DQ random charts of 10% of submitted cases. The CathPCI Registry audit focuses on variables used for the existing PCI mortality models.

Finally, we assess the reliability of the data elements by comparing model variable frequencies and odds ratios in two years of data.

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 asses 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, then measured again using a second random subset exclusive of the first, and finally comparing the agreement between the two resulting performance measures across hospitals (Rousson et al., 2002).

For test-retest reliability, we combined index admissions from successive measurement periods into one dataset (2010 and 2011), 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 (ICC) (Shrout and Fleiss, 1979), and assessed the values according to conventional standards (Landis and Koch, 1977). Specifically, we used the two data samples and calculated the risk-standardized readmission rate (RSRR) for each hospital for each sample. The agreement of the two RSRRs was quantified for hospitals in each sample using the intra-class correlation (ICC) as defined by Shrout and Fleiss (1979).

Using two independent samples provides an honest 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 small volume hospitals contribute less ´signal´. As such a split sample using a single measurement period likely introduces extra noise; potentially underestimating the actual test-retest reliability that would be achieved if the measures were reported using additional years of data. Furthermore, the measure is specified for the entire PCI population, but we tested it only in the subset of Medicare FFS patients for whom information about vital status was available. This reduced the cohort available for testing by approximately 40%.

References:

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

2) Shrout P, Fleiss J. Intraclass correlations: uses in assessing rater reliability. Psychological Bulletin 1979;86:420-428.

3) 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

Overall, risk factor frequencies changed little across years, and there were no notable differences in the odds ratios across years of data (Table 1). See the 2013 Measure Updates and Specifications Report in the attached Appendix for details.

Table . 30-Day Readmission Model (GLM) Risk Factor Frequency by Year of Discharge (%)

| **Description** | **2006**  **(Validation)**  **N=117,375 in 618 Hospitals** | **2007 (Development)**  **N=128,745 in 766 Hospitals** | **2010-2011**  **N=277,512 in 1,197 Hospitals** |
| --- | --- | --- | --- |
| Age/10 (SD) | 74.7 (6.5) | 74.7 (6.6) | 75.12 (6.9) |
| Female | 41.8 | 41.2 | 39.9 |
| BMI/5 | | |  |
| Unknown | 0.1 | 0.1 | 0.2 |
| Mean (SD) | 28.5 (5.7) | 28.6 (5.8) | (5.4) |
| Heart failure - previous history | 13.8 | 13.8 | 16.6 |
| Previous valvular surgery | 1.6 | 1.7 | 2.2 |
| Cerebrovascular Disease | 16.0 | 16.0 | 17.8 |
| Peripheral Vascular Disease | 15.6 | 15.6 | 16.8 |
| Chronic Lung Disease | 18.6 | 18.6 | 18.9 |
| Diabetes | | |  |
| Non-Insulin diabetes | 22.4 | 22.6 | 22.9 |
| Insulin diabetes | 9.8 | 10.1 | 13.1 |
| Glomerular Filtration Rate (GFR) | | |  |
| GFR: 0=Not measured | 4.0 | 3.7 | 5.7 |
| GFR: 1="0<=GFR<30" | 4.0 | 4.3 | 5.1 |
| GFR: 2="30<=GFR<60" | 36.6 | 37.2 | 32.8 |
| GFR: 4="GFR>=90" | 8.3 | 8.3 | 11.8 |
| Renal Failure - Dialysis | 1.6 | 1.9 | 2.7 |
| Hypertension | 81.8 | 82.9 | 86.9 |
| History of Tobacco Use | 11.8 | 11.9 | 13.6 |
| Previous PCI | 35.9 | 37.2 | 40.7 |
| Heart failure - current status | 12.0 | 11.9 | 13.4 |
| Symptoms present on admission | | |  |
| No MI on admission | 75.4 | 73.5 | 64.6 |
| MI after 24 hours on  admission | 5.7 | 6.0 | 2.5 |
| Ejection Fraction (EF) Percentage | | |  |
| EFP: 1=Not measured | 28.3 | 28.5 | 29.7 |
| EFP: 2="0<=EFP<30" | 3.9 | 3.9 | 4.6 |
| EFP: 3="30<=EFP<45" | 11.9 | 11.9 | 12.1 |
| PCI status | | |  |
| PCI status: 2=Urgent | 36.0 | 36.4 | 44.4 |
| PCI status: 3=Emergency | 11.1 | 12.2 | 14.9 |
| PCI status: 4=Salvage | 0.1 | 0.1 | 0.1 |
| Highest Risk Lesion – location | | |  |
| pRCA/mLAD/pCIRC | 38.2 | 37.9 | 37.9 |
| pLAD | 17.6 | 17.3 | 16.9 |
| Left main | 2.4 | 2.4 | 3.2 |
| Highest Pre-Procedure TIMI Flow: None | 7.8 | 8.7 | 11.3 |

Measure score reliability results

In the most recent years of data (2010-2011), there were 277,512 admissions in the combined two-year sample, with 138,756 admissions to 1,190 hospitals in the first randomly selected sample (mean RSRR 12.5%), and 138,756 admissions to 1,193 hospitals in the second randomly-selected sample (mean RSRR 12.1%). The agreement between the two RSRRs for each hospital was 0.3711, which according to the conventional interpretation is “fair” (Landis & Koch, 1977). The intra-class correlation coefficient is based on a split sample of 2 years of data, resulting in a volume of patients in each sample equivalent to only 1 year of data, whereas the measure is likely to be reported with a full two years of data

Reference.

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

**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 indicate that the underlying data elements are reliable. Additionally, the ICC score demonstrates fair agreement across samples using a “strict” approach to assessment that would likely improve with greater sample size.

References

Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. Mar 1977;33(1):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 other measures, through use of established measure development guidelines, by systematic assessment of measure face validity by a technical expert panel (TEP) of national experts and stakeholder organizations, and through registry data validation.

Validity of Registry Data

Data element validity testing was done on the specified measure by comparing with variables in the ACC audit program. The NCDR CathPCI Registry has an established DQP that serves to assess and improve the quality of the data submitted to the registry. There are two complementary components to the Data Quality Program- the Data Quality Report (DQR) and the Data Audit Program (DAP). The DQR process assesses the completeness of the electronic data submitted by participating hospitals. Hospitals must achieve >95% completeness of specific data elements identified as “core fields” to be included in the registry’s data warehouse for analysis. The “core fields” encompass the variables included in our risk adjustment models. The process is iterative, providing hospitals with the opportunity to correct errors and resubmit data for review and acceptance into the data warehouse. All data for this analysis passed the DQR completeness thresholds.

The DAP consists of annual on-site chart review and data abstraction. Among participating hospitals that pass the DQR for a minimum of two quarters, at least 5% are randomly selected to participate in the DAP. At individual sites, auditors review charts of 10% of submitted cases. The audits focus on variables that are used in the NCDR risk-adjusted in-hospital mortality model including demographics, comorbidities, cardiac status, coronary anatomy, and PCI status. However, the scope of the audit could be expanded to include additional fields. The DAP includes an appeals process for hospitals to dispute the audit findings. The NCDR DAP was accepted by the National Quality Forum as part of its endorsement of the CathPCI Registry’s in-hospital risk-adjusted mortality measure.

Additionally, we compared the model performance in the development sample with its performance in a similarly derived sample from patients discharged in 2006 who had undergone PCI. There were 117,375 cases discharged from the 618 hospitals in the 2006 validation dataset. This validation sample had a crude readmission rate of 10.7%. The performance was not substantively different in this validation sample (ROC=0.663), as compared to the development sample (ROC=0.665). As the results in Table 2 show, the 2006 and 2007 models are similarly calibrated.

We also examined the temporal variation of the standardized estimates and frequencies of the variables in the development and validation models.

To assess the predictive ability of the model, we grouped patients into deciles of predicted 30-day readmission and compared predicted readmission with observed readmission for each decile in the derivation cohort.

To evaluate model performance after the re-specification to Version 4 variables, we compared the odds ratios (OR) and c-statistics in 2008 Version 3 data and 2010 Version 4 data.

Validity as Assessed by External Groups

During original measure development and in alignment with the CMS Measures Management System (MMS), we released a public call for nominations and convened a TEP when originally developing the measure. The purpose of convening the TEP was to provide input and feedback during measure development from a group of recognized experts in relevant fields. The TEP represented physician, consumer, hospital, and purchaser perspectives, chosen to represent a diverse of perspectives and backgrounds. Please see the Appendix attachment for details.

ICD-9 to ICD-10 Conversion

Statement of Intent

[X] Goal was to convert this measure to a new code set, fully consistent with the intent of the original measure.

[ ] Goal was to take advantage of the more specific code set to form a new version of the measure, but fully consistent with the original intent.

[ ] The intent of the measure has changed.

*Process of Conversion*

ICD-10 codes were initially identified using 2013 General Equivalence Mapping (GEM) 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 or Code Table).

*Lead clinical expert*

Jeptha Curtis, M.D., Associate Professor of Medicine, Section of Cardiovascular Disease, Yale University

Table . 30-Day Readmission Model Performance: Results Based on the GLM

|  |  |  |
| --- | --- | --- |
| **Indices** | **Development Sample** | **Validation Sample** |
| Year | 2007 | 2006 |
| N | 128745 | 117375 |
| RR | 11.1% | 10.7% |
| Calibration (γ0, γ1)1 | (0.00, 1.00) | (-0.06, 0.99) |
| Discrimination- Adjusted R-Square2 | 0.07 | 0.06 |
| Discrimination -Predictive Ability3 (lowest decile %, highest decile %) | (4.05, 25.08) | (3.80, 23.80) |
| Discrimination – ROC | 0.665 | 0.663 |
| Residuals Lack of Fit (Pearson Residual Fall %) |  |  |
| <-2 | 0.00 | 0.00 |
| [-2, 0) | 88.86 | 89.33 |
| [0, 2) | 2.21 | 1.85 |
| [2+ | 8.93 | 8.82 |
| Model χ2 [Number of Covariates]4 | 4448.36 [31] | 3812.62 [31] |

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

The performance of the development and validation samples is similar. The areas under the receiver operating characteristic (ROC) curve are 0.665 and 0.663, respectively, for the two samples. In addition, they are similar with respect to predictive ability. For the development sample, the predicted readmission rate ranges from 4% in the lowest predicted decile to 25% in the highest predicted decile, a range of 21%. For the validation sample, the corresponding range is 4% to 24%, a range of 20%.

Additionally, the frequencies and regression coefficients are fairly consistent over the two years of data. Also, there was excellent correlation between predicted and observed readmission.

We estimated hospital-level RSRRs using the corresponding hierarchical logistic regression samples for the linked patient sample with cases performed in 2010-2011. We then examined the linear relationship between the two sets of estimates using regression techniques and weighting by the total number of cases in each hospital. The correlation coefficient of the standardized rates from the administrative and medical record models is 0.999.

The c-statistic for the 2010, Version 4 model was 0.680. This is a negligible change from the 2008, Version 3 model, which had a c-statistic of 0.676. Odds ratios in both data years are comparable, further indicating that model performance was not significantly altered by re-specification to Version 4 variables. The current model can use the Version 4 registry data.

**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 audits conducted by the ACC support the overall validity of the data elements included in this measure. The data elements used for risk adjustment were consistently found for all patients and were accurately extracted from the medical record.

Additionally, theresults between the development and validation samples proved to be similar in each of the model testing that was performed. The ROC results were nearly identical. The correlation between the resulting RSRRs calculated from both models was 0.999 which demonstrates observed readmission is similar to predicted readmission.

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

**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*)

All exclusions were determined by careful clinical review and have been made based on clinically relevant decisions. To ascertain impact of exclusions on the cohort, we examined overall frequencies and proportions of the total cohort excluded for each exclusion criterion. These exclusions are consistent with similar NQF-endorsed readmission measures. Rationales for the exclusions are detailed in data field S.10 (Denominator Exclusions).

**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*)

We examined overall frequencies and proportions of the admissions excluded for each exclusion criterion in the most recent data (2010-2011). The initial sample without exclusions included 395,665 admissions to 1,203 hospitals. After applying the exclusion criteria as outlined in Table 3, the 2010-2011 study sample included 277,212 patients admitted to 1,197 hospitals.

Table . Exclusions from the target population for the combined 2010-2011 study sample.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Exclusions** | **2010-2011** | | | | | | | | |
| **Patient Stay** | | | | **Hospitals** | | | | |
| **# (%)** | | | | **# (%)** | | | | |
| **Initial Sample** | **395,665** | | | | **1,203** | | | | |
| Not Medicare Fee-for-Service (FFS) patient on admission | 88,509 (22.37) | | | | 2 (0.17) | | | | |
| **Remaining** | **307,156** | | | | **1,201** | | | | |
| Not the first claim in the same claim bundle\* | 11 (0.00) | | | | 0 (0.00) | | | | |
| **Remaining** | **307,145** | | | | **1,201** | | | | |
| Get the procedure more than 10 days after admission | 2,286 (0.74) | | | | 0 (0.00) | | | | |
| **Remaining** | **304,859** | | | | **1,201** | | | | |
| Transferred out | 2,102 (0.69) | | | | 1 (0.08) | | | | |
| **Remaining** | **302,757** | | | | **1,200** | | | | |
| In-hospital death | 6,909 (2.28) | | | | 2 (0.17) | | | | |
| **Remaining** | **295,848** | | | | **1,198** | | | | |
| Against medical advice | 421 (0.14) | | | | 0 (0.00) | | | | |
| **Remaining** | **295,427** | | | | **1,198** | | | | |
| Not a full month follow-up | 10,943 (3.70) | | | | 1 (0.08) | | | | |
| **Remaining** | **284,484** | | | | **1,197** | | | | |
| Duplicate admissions\*\* | 6,972 (2.45) | | | | 0 (0.00) | | | | |
| **Study Sample** | **277,512** | | | | **1,197** | | | | |
| Readmission within 30-days\*\*\* | 39,078 (14.08) | | | |  | | | | |
| Readmission within 30-days\*\*\*\* | 32,601(11.7) | | | |  | | | | |
| \* Defined as two or more claims in which the admission date of the current claim is before or the same as the discharge date of its previous claim. When this happens, the information at discharge of the first claim are replaced by the information at discharge of the last claim. | | | | | | | | | | |
| \*\* Defined as admissions within 30 days of a PCI index admission. | | |  |  |  | |  |  |  |  | |  |  |
| \*\*\* Same claim are not considered as readmission: Same Hospital, Same Principle Diagnosis, Same Group of Procedure (PCI), SAME day (admitted within one day of previous discharge). | | | | | | | | | | |
| \*\*\*\* Planned readmissions and unplanned readmission following a planned readmission were not considered as readmission. | | | | | | | | | | |

**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 majority of exclusions are necessary to 1) link registry and administrative data (e.g. excluding patient not enrolled in Medicare FFS) and 2) identify patients eligible for readmission (e.g. excluding patients who died before discharge). As such, these exclusions are not discretionary and do not require further testing. There are two exclusions that are discretionary but both are infrequent. Only 0.14% of patients left against medical advice, and only 0.74% of patients had their PCI performed more than 10 days after hospital admission. None of the exclusions resulted in any unexpected shifts in the performance 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** 20 **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*)

We sought to develop a model that included key variables that were clinically relevant and based on strong association with 30-day readmission.

To create a model with increased usability while retaining excellent model performance, we tested the performance of the model without those variables considered to be questionably feasible. To select candidate variables, a team of clinicians reviewed all variables in the NCDR CathPCI Registry database (a copy of the data collection form and the complete list of variables collected and submitted by hospitals can be found at www.ncdr.com). We did not consider as candidate variables those that we would not want to adjust for in a quality measure, such as potential complications, certain patient demographics (e.g., race, socioeconomic status), and patients‟ admission path (e.g., admitted from a skilled nursing facility [SNF]).

Based on careful clinical review and further informed by a review of the literature, a total of 29 variables were determined to be appropriate for consideration as candidate variables (Table 4).

For categorical variables with missing values, the value from the reference group was added. The percentage of missing values for all categorical variables was very small (<1%). There were three continuous variables with missing values: body mass index (BMI, 0.1%), glomerular filtration rate (GFR, 3.7%), and left ventricular ejection fraction (LVEF, 28.5%); we considered the missing of GFR and LVEF as an independent category of “unmeasured” and for BMI; we stratified by gender and imputed the missing values to the median of the corresponding groups.

We used logistic regression with stepwise selection (entry p<0.05; retention with p<0.01) for variable selection. We also assessed the direction and magnitude of the regression coefficients. This resulted in a final risk-adjusted readmission model that included 20 variables (Table 5). There were variables for demographics (age and gender), history and risk factors, cardiac status (heart failure, symptoms present on admission), cath lab visits (ejection fraction percentage), and PCI procedure (PCI status, highest risk lesion, highest pre-procedure TIMI flow).

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

Table . PCI Readmission Candidate Variables

| **Description** | **NCDR Item Number** | **Code** |
| --- | --- | --- |
| **Demographic** | | |
| Age | 252 | Age |
| Female | 260 | FEMALE |
| **History and Risk Factors** | | |
| BMI | Derived (410, 412) | BMI |
| Previous MI | 420 | PrevMI |
| Heart Failure-previous history | 424 | PrCHF |
| Previous valvular surgery | 426 | PrValve |
| Cerebrovascular Disease | 450 | CVD |
| Peripheral Vascular Disease | 452 | PVD |
| Chronic Lung Disease | 454 | CLD |
| Diabetes | Derived (430, 432) | NewDIAB |
| None | | Reference |
| Non-insulin diabetes | | NEWDIAB1 |
| Insulin diabetes | | NEWDIAB2 |
| Glomerular Filtration Rate (GFR) | Derived (252, 260, 270, 439, 440) | GFR |
| Not measured | Derived | GFRGRP0 |
| GFR<30 | Derived | GFRGRP1 |
| 30≤GFR<60 | Derived | GFRGRP2 |
| 60≤GFR<90 | | Reference |
| GFR≥90 | Derived | GFRGRP4 |
| Renal failure-dialysis | 444 | Dialysis |
| Hypertension | 456 | Hypertn |
| History of tobacco use | 460 | Tobacco |
| Family history of CAD | 480 | FHCAD |
| Previous PCI | 490 | PrPCI |
| Previous CABG | 494 | PrCAB |
| **Cardiac Status** | | |
| Heart failure - current status | 500 | CHF |
| NYHA | 510 | ClassNYH |
| Class I or II | | Reference |
| Class III | Derived | NYHC3 |
| Class IV | Derived | NYHC4 |
| Cardiogenic shock | | 520 |
| ST elevation MI (STEMI) | Derived (550, 560, 812) | STEMI |
| Symptoms present on admission | Derived (550, 560) | AdmSxPre |
| No MI | | ADMSX1 |
| MI within 24 hours | | Reference |
| MI after 24 hours | | ADMSX3 |
| Cath Lab Visit | | |
| Ejection Fraction (EF) Percentage | Derived (654, 656) | HDEFGRP |
| Not measured | | HDEFGRP1 |
| EF<30 | | HDEFGRP2 |
| 30≤EF<45 | | HDEFGRP3 |
| EF≥45 | | Reference |
| Left main disease | Derived (660, 661) | LMGT50 |
| Number of vessels with disease | Derived (662 to 671) | VESSELD |
| ≤1 | Reference | |
| 2 | Derived | VESSELD2 |
| 3 | Derived | VESSELD3 |
| **PCI Procedure** | | |
| PCI status | 804 | PCIStat |
| Elective | Reference | |
| Urgent | Derived | PCIS2 |
| Emergency | Derived | PCIS3 |
| Salvage | Derived | PCIS4 |
| Highest Lesion location | Derived (900, 902) | NLESLOC |
| pRCA/mLAD/pCIRC | Derived | NLESLOC1 |
| pLAD | Derived | NLESLOC2 |
| Left main | Derived | NLESLOC3 |
| Other | Derived | |
| Highest pre-procedure TIMI∗∗flow: none | 920 | NPRETIMI |
| Highest risk lesion: SCAI∗∗∗ \_lesion class | Derived (910, 950) | NSCAILC |
| I | Reference | |
| II | Derived | NSCAILC2 |
| III | Derived | NSCAILC3 |

Table . PCI Readmission Final Model Variables

| **Description**  (n) Variable | **Code** |
| --- | --- |
| **Demographic** | |
| 1. Age | Age |
| (2) Female | FEMALE |
| **History and Risk Factors** | |
| (3) Body Mass Index | BMI |
| (4) Heart failure-previous history | PRCHF |
| (5) Previous valvular surgery | PRVALVE |
| (6) Cerebrovascular Disease | CVD |
| (7) Peripheral Vascular Disease | PVD |
| (8) Chronic Lung Disease | CLD |
| (9) Diabetes |  |
| None | Reference |
| Non-insulin diabetes | NEWDIAB1 |
| Insulin diabetes | NEWDIAB2 |
| (10) Glomerular Filtration Rate (GFR) |  |
| Not measured | GFRGRP0 |
| GFR<30 | GFRGRP1 |
| 30≤GFR<60 | GFRGRP2 |
| 60≤GFR<90 | Reference |
| GFR≥90 | GFRGRP4 |
| (11) Renal failure - dialysis | DIALYSIS |
| (12) Hypertension | HYPERTN |
| (13) History of tobacco use | TOBACCO |
| (14) Previous PCI | PrPCI |
| **Cardiac Status** | |
| (15) Heart failure – current status | CHF |
| (16) Symptoms present on admission |  |
| No MI | ADMSX1 |
| MI within 24 hours | Reference |
| MI after 24 hours | ADMSX3 |
| **Cath Lab Visit** | |
| (17) Ejection Fraction (EF) Percentage |  |
| Not measured | HDEFGRP1 |
| EF<30 | HDEFGRP2 |
| 30≤EF<45 | HDEFGRP3 |
| EF≥45 | Reference |
| **PCI Procedure** | |
| (18) PCI status | |
| Elective | Reference |
| Urgent | PCIS2 |
| Emergency | PCIS3 |
| Salvage | PCIS4 |
| (19) Highest risk lesion – location |  |
| pRCA/mLAD/pCIRC | NLESLOC1 |
| pLAD | NLESLOC2 |
| Left main | NLESLOC3 |
| Other | Reference |
| (20) Highest pre-procedure TIMI flow: none | |

**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*)

Approach to assessing model performance

During measure development, we computed three summary statistics for assessing model performance (Harrell and Shih, 2001) for the development and validation cohort:

***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 the development dataset but fails to provide valid predictions in new patients)

We compared the model performance in the development sample with its performance in a similarly derived sample from patients discharged in 2006 who had undergone PCI. There were 117,375 cases discharged from the 618 hospitals in the 2006 validation dataset. This validation sample had a crude readmission rate of 10.7%. We also computed statistics (1) and (2) for the current measure cohort, which includes discharges from 2010-2011.

*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*)**:**

For the development cohort the results are summarized below:

C-statistic=0.665

Predictive ability (lowest decile %, highest decile %): 4.05%, 25.08%

For the validation cohort the results are summarized below:

C statistic=0.663

Predictive ability (lowest decile %, highest decile %): 3.80%, 23.80%

For the current measure cohort (combined data from 2010 and 2011) the results are summarized below:

C statistic=0.668

Predictive ability (lowest decile %, highest decile %): 4.2%, 26.1%

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

For the development cohort the results are summarized below:

Calibration: (0.00,1.00)

For the validation cohort the results are summarized below:

Calibration: (-0.06, 0.99)

For the current measure cohort the results are summarized below:

Calibration: (-0.004, 1.008)

**2b4.8. Statistical Risk Model Calibration – Risk 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 showing the distributions for the current measure cohort.

Figure . Risk decile plot, 2010-2011 study sample.

**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*)

**Discrimination Statistics**

The C-statistics of 0.665, 0.663 and 0.668 indicate good model discrimination. Readmission, as opposed to other outcomes such as mortality consistently has a lower c-statistic, even in medical record models. This is likely because readmission is less determined by patient comorbidities and more by health system factors. The model indicated a wide range between the lowest decile and highest decile, indicating the ability to distinguish high-risk patients from low-risk patients.

**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 show a good calibration of the model. This 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)*

For the currently publicly reported measures of hospital outcomes, including the PCI readmission measure, CMS estimates an interval estimate for each risk-standardized rate to characterize the amount of uncertainty associated with the rate. It then 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 U.S. national rate (NCDR registry rate for PCI). We assessed variation in RSRRs among hospitals by examining the distribution of the hospital RSRRs and plotting the histogram of the hospital RSRRs.

**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*)

Recent analyses of Medicare FFS data show variation in RSRRs among hospitals. Using the most recent data sample (2010-2011) and updating the measure by applying Version 2.1 of the planned readmission algorithm, the mean hospital RSRR was 11.8%, with a range of 8.5% to 16.7%. The interquartile range was 10.9% to 12.6%. Out of 361 hospitals voluntarily reporting rates on Hospital Compare in 2013, 14 performed “better than the NCDR registry average,” 297 performed “no different from the NCDR registry average,” and 2 performed “worse than the NCDR registry average.” Additionally, 31 did not have results available for the reporting period. 17 were classified as “number of cases too small” (fewer than 25) to reliably tell how well the hospital is performing.

Figure 3. Distribution of risk-standardized mortality rates (RSMRs); 2010-2011 combined sample.



**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 variation in rates suggests there are clinically meaningful differences across hospitals in the 30-day risk-standardized readmission after PCI.

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

**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 . This measure has only one set of specifications.

**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 . This measure has only one set of specifications.

**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 . This measure has only one set of specifications.

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

**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*)

We examined rates of missing data for all candidate variables and examined histograms of the frequency of missingness by hospital.

**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*)

Overall, the percentage of missing values for all categorical variables was very small (<1%). There were three continuous variables with missing values: body mass index (BMI, 0.1%), glomerular filtration rate (GFR, 3.7%), and left ventricular ejection fraction (LVEF, 28.5%); we considered the missing of GFR and LVEF as an independent category of “unmeasured” and for BMI; we stratified by gender and imputed the missing values to the median of the corresponding groups.

**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*)

As noted above, model performance was comparable when we included or excluded cases with missing data.