**National Quality Forum—Measure Testing (subcriteria 2a2, 2b2-2b6)**

**Measure Title**:

Risk-Adjusted Coronary Artery Bypass Graft (CABG) Readmission Rate

**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 specifications** (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.  **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.    **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**.  **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*).

STS Adult Cardiac Surgery Database and Medicare Part A Claims

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

2008 – 2010

**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*)

Analysis #1: Data abstraction reliability

Data abstraction reliability was assessed by the STS national audit program for the year 2011. A total of 51 STS sites were audited in 2011. See “analytic methods” for details.

Analysis #2: Statistical reliability (signal-to-noise ratio)

Statistical reliability of the proposed readmission measure was assessed using linked STS-CMS data for calendar years 2008-2010. The analysis population consisted of Medicare fee-for-service beneficiaries age 65 years or older at discharge who underwent isolated coronary artery bypass grafting (CABG) during 2008-2010 and were discharged alive. Candidate CABG admissions were identified by selecting Medicare Part A claims with an ICD-9-CM procedural code for CABG (36.1x) in any position. Records were retained for analysis if they met the following additional criteria:

1. Linked to an STS record for isolated CABG (see below for record linkage criteria and definition of isolated CABG);
2. Eligible for Medicare fee-for-service (FFS) A and B for at least two months after discharge or until month of death, whichever is first;
3. Discharged from acute care setting within 1 year of index CABG admission;
4. Did not leave against medical advice;
5. No logically inconsistent claims data (e.g. claims with overlapping admission and discharge dates);
6. Is the first eligible operation per patient during the measurement period

After applying these inclusion/exclusion criteria, the analysis population included 162,503 index CABG admissions from 1010 CMS hospitals. Results are reported for the subset of hospitals for which at least 90% of eligible CMS records could be linked to a corresponding STS record (150,820 index admissions, 900 CMS hospitals).

**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*)

Please see response in 1.5 above

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

N/A

**2a2. RELIABILITY TESTING**

***Note****: If accuracy/correctness (validity) of data elements was empirically tested*, *separate reliability testing of data elements is not required – in 2a2.1 check critical data elements; in 2a2.2 enter “see section 2b2 for validity testing of data elements”; and skip 2a2.3 and 2a2.4.*

**2a2.1. What level of reliability testing was conducted**? (*may be one or both levels*)

☒ **Critical data elements used in the measure** (*e.g., inter-abstractor reliability; data element reliability must address ALL critical data elements*)   
☒ **Performance measure score** (e.g., *signal-to-noise analysis*)

**2a2.2. For each level checked above, describe the method of reliability testing and what it tests** (*describe the steps―do not just name a method; what type of error does it test; what statistical analysis was used*)

Analysis #1: Data abstraction reliability

In 2006, STS initiated an independent audit program to evaluate the accuracy and completeness of the STS-ACSD and to improve data collection protocols and processes. Each year, 5% of STS-ACSD sites are randomly selected to be audited by an independent auditing firm. The audit process involves re-abstraction of data for 20 cases and comparison of 75 designated data elements with those submitted to the data warehouse. In addition, a surgical log comparison is performed to ensure all appropriate records are submitted.

Analysis #2: Statistical reliability (signal-to-noise ratio)

Definition of Reliability –

Reliability is conventionally defined as the proportion of variation in a performance measure that is due to true between-hospital differences (i.e., signal) as opposed to random statistical fluctuations (i.e., noise). A mathematically equivalent definition is the squared correlation between a measurement and the true value. This quantity cannot be calculated directly but may be estimated in a model. To do this, we implemented a fully Bayesian version of our proposed hierarchical model and used it to estimate hospital-specific risk-standardized readmission rates (HSRR’s) as defined elsewhere in our measure documentation. We also performed an unadjusted version to model variation in hospital-specific readmission rates without adjusting for case mix.

**Technical Details**

Let denote the true unknown readmission rate for the *j*-th of *J* hospitals. Prior to estimating reliability, the numerical value of was estimated for each hospital. Estimation was done using Markov Chain Monte Carlo (MCMC) simulations and involved the following steps:



First, for each *j*, we randomly generated a large number *N* of possible numerical values of by sampling from the Bayesian posterior probability distribution of . Let denote the *i*-th of these *N* randomly sampled numerical values for the *j*-th hospital.

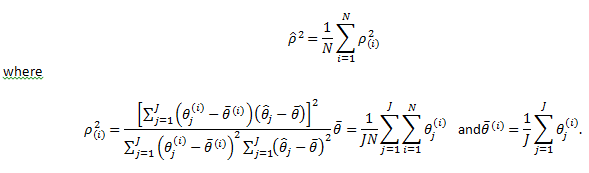


Second, for each *j*, a Bayesian estimate of was calculated as the arithmetic average of the randomly sampled values; in other words .



Our reliability measure was defined as the estimated squared correlation between the set of hospital-specific estimates and the corresponding unknown true values . Let denote the unknown true squared correlation of interest and let denote an estimate of this quantity. The estimate was calculated as





A 95% Bayesian probability interval for was obtained calculating the 2.5th and 97.5th percentiles of the set of numbers …,. An analogous calculation was used to estimate reliability for readmission rates not adjusted for case mix.



**2a2.3. For each level 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*)

Analysis #1: Data abstraction reliability

The results from the 2011 audit revealed an overall agreement rate of 96.32% for audited data elements. The case log submission agreement was 99.61%. Mismatches were typically related to unscheduled emergency cases.

Analysis #2: Statistical reliability (signal-to-noise ratio)

We compared the reliability of risk-standardized readmission rates (RSRR’s) estimated with one year and three years of data.

**Estimated Model Reliability by Estimated Correlation Statistic with 95% Credible Intervals, Over All Three Years in Hospitals with at Least 10, 30, 50, 100, 200, and 300 Cases in Three Years**



|  |  |  |
| --- | --- | --- |
| **Minimum Number of Included Cases for Hospital Inclusion** | **Unadjusted Model Reliability** | **Adjusted Model Reliability** |
| 10 | 0.493 (0.447, 0.536) | 0.455 (0.409, 0.499) |
| 30 | 0.510 (0.463, 0.553) | 0.470 (0.424, 0.515) |
| 50 | 0.533 (0.486, 0.577) | 0.492 (0.444, 0.537) |
| 100 | 0.582 (0.533, 0.627) | 0.543 (0.493, 0.591) |
| 200 | 0.686 (0.630, 0.739) | 0.641 (0.582, 0.699) |
| 300 | 0.786 (0.723, 0.841) | 0.714 (0.639, 0.781) |

**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?*)

Approximately 44% of variation in RSRR’s is explained by true signal variation as opposed to random statistical fluctuations. This level of reliability is comparable or higher than other CMS readmission measures.

**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 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)*

Please refer to the reliability testing section above for information about the accuracy of STS data collection and to the risk adjustment section below for information about the validity of the proposed risk adjustment methodology.

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

Please refer to the reliability testing section above for information about the accuracy of STS data collection and to the risk adjustment section below for information about the validity of the proposed risk adjustment methodology.

**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?*)

Please refer to the reliability testing section above for information about the accuracy of STS data collection and to the risk adjustment section below for information about the validity of the proposed risk adjustment methodology.

**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*)

Please refer to section S.9 and S.10 for information on 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*)  
Please refer to section S.9 and S.10 for information on exclusions

**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*)

Please refer to section S.9 and S.10 for information on exclusions

**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** 21 covariates **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

**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*)

Risk adjustment procedures were developed and tested using linked STS-CMS data from 2008-2010. The analysis population consisted of Medicare fee-for-service beneficiaries age 65 years or older at discharge who underwent isolated coronary artery bypass grafting (CABG) during 2008-2010 and were discharged alive. Candidate CABG admissions were identified by selecting Medicare Part A claims with an ICD-9-CM procedural code for CABG (36.1x) in any position. Records were retained for analysis if they met the following additional criteria:

1. Linked to an STS record for isolated CABG (see below for record linkage criteria and definition of isolated CABG);
2. Eligible for Medicare fee-for-service (FFS) A and B for at least two months after discharge or until month of death, whichever is first;
3. Discharged from acute care setting within 1 year of index CABG admission;
4. Did not leave against medical advice;
5. No logically inconsistent claims data (e.g. claims with overlapping admission and discharge dates);
6. Is the first eligible operation per patient during the measurement period

Candidate covariates

Candidate covariates and their functional form were identified from the published STS 2008 isolated CABG mortality and morbidity models (Shahian et al.). Race and ethnicity were candidate covariates in the STS 2008 CABG models but were not candidate covariates for the current readmission measure project.

Selection of final model covariates

Final model covariates were selected by an expert surgeon panel after performing a variety of exploratory analyses to help inform the surgeon panel’s decision. First, for each of the 2008, 2009, and 2010 datasets separately, we performed the following sets of analyses:

1. Ordinary logistic regression with stepwise variable selection.

* Significance level to enter = 0.05, significance level to stay = 0.05
* Significance level to enter = 0.01, significance level to stay = 0.01

1. Bootstrap analysis of ordinary logistic regression with stepwise variable selection.

* Significance level to enter = 0.05, significance level to stay = 0.05.
* 1000 bootstrap samples were generated by sampling with replacement from the full dataset.
* Backward selection was applied to each bootstrap sample independently.
* For each candidate covariate, we recorded the number of times (out of 1000) that the covariate was selected into the final stepwise model. More important variables will be more likely to be selected in the final models.

The main goal of these analyses was to identify important covariates (i.e. those with large odds ratios) and to assess the consistency of variable selection across calendar years and within bootstrap samples from the same calendar year. A secondary purpose was to explore the functional form of continuous variables and to assess whether categorical variables with several categories could be collapsed into a smaller number of categories. Categorical variables were parameterized such that removing a variable caused 2 adjacent categories to be collapsed into a single category. Continuous variables were modeled as piecewise linear functions and parametrized such that removing a variable caused two adjacent line segments (2 slopes) to be collapsed into a single line segment (a single slope).

The final model selected by the surgeon panel included all covariates that were either (1) selected at the 0.05 level in the original full sample for at least one calendar year; or (2) were selected in at least 50% of bootstrap replicates at the 0.05 level in at least one calendar year. Based on these covariates were selected for the final model:

1. Ejection Fraction
2. Preoperative Atrial Fibrillation
3. Unstable Angina (no MI <= 7 days)
4. Myocardial Infarction
5. Age
6. Congestive Heart Failure
7. Renal Function
8. Status
9. Gender
10. Reoperation
11. Chronic Lung Disease
12. Diabetes
13. Preoperative IAPB or Inotrope
14. Immunosuppressive Treatment
15. PVD
16. Body Surface Area
17. CVD
18. Hypertension
19. PCI <= 6 hours
20. Left Main Disease
21. Surgery Date

After selecting final model covariates, the final model was re-estimated using data from 2008 and tested in data from 2009. In addition, the final model was re-estimated using the entire 2008-2010 dataset. Finally, the model was re-estimated in the 2008-2010 sample using hierarchical instead of ordinary logistic regression. The hierarchical version of the model was used for calculating hospital-specific risk-standardized readmission rates (RSRR’s) as described below.

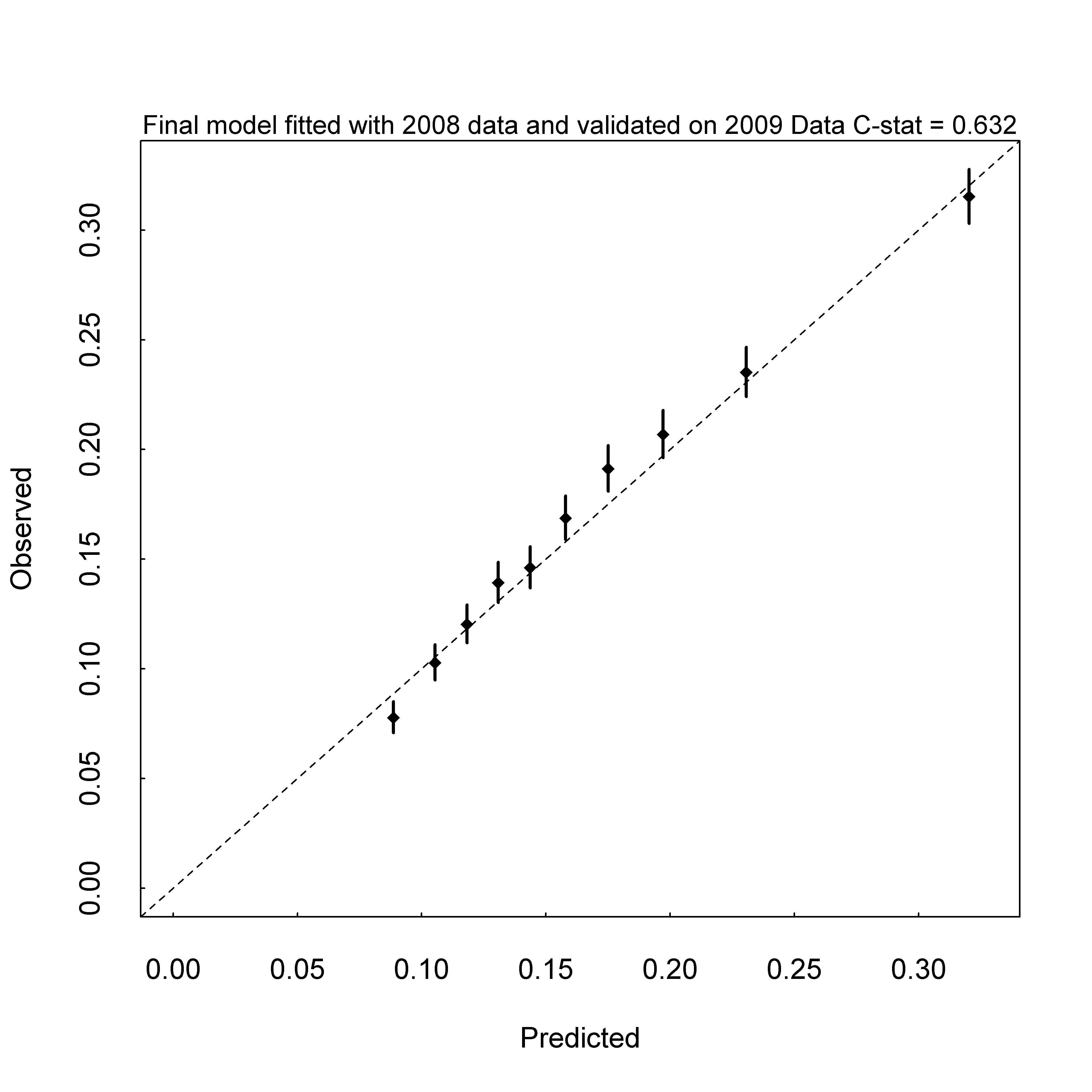
Calculation of risk-standardized readmission rates (RSRR’s)

The methodology of calculating RSRR’s was chosen to be 100% consistent with other existing CMS readmission measures. Briefly, we used hierarchical logistic regression with hospital-specific random intercept parameters. The following description was copied from documentation produced by Yale University for the previous CMS PCI readmission measure: “These rates are calculated as the ratio of the predicted number of readmissions to the expected number of readmissions, multiplied by the national unadjusted readmission rate. The expected number of readmissions for each hospital was estimated using its patient mix and the average hospital-specific intercept. The predicted number of readmissions in each hospital was estimated given the same patient mix but an estimated hospital-specific intercept. Operationally, the expected number of readmissions for each hospital is obtained by summing the expected readmission rates for all patients in the hospital. The expected readmission rate for each patient is calculated via the hierarchical model by applying the subsequent estimated regression coefficients to the observed patient characteristics and adding the average of the hospital-specific intercepts. The predicted number of readmissions for each hospital is calculated by summing the predicted readmission rates for all patients in the hospital. The predicted readmission rate for each patient is calculated through the hierarchical model by applying the estimated regression coefficients to the patient characteristics observed and adding the hospital-specific intercept.” Our method of calculating 95% interval estimates for RSRR’s was also identical to previous CMS readmission measures developed by Yale University.

Testing Results

After selecting the final list of model covariates, coefficients were re-estimated using data from 2008 only. Calibration and discrimination were then assessed in a separate sample of 2009 data. To assess calibration, we compared observed versus expected all-cause 30-day readmission rates within subgroups based on deciles of predicted risk in 2009 data. Discrimination was assessed by the C-index (area under the receiver operating characteristics [ROC] curve). Predictive ability was further assessed by comparing the average predicted risk in the lowest-risk versus highest-risk decile of predictive risk.

* **Calibration**: As shown in the figure and table below, there was excellent agreement between observed and expected readmission rates.
* **Discrimination**: The C-index was 0.632. This level of discrimination is comparable to several other existing CMS readmission measures.
* **Predictive ability**: Predicted risk ranged from 7.8% in the lowest-risk decile to 31.5% in the highest-risk decile.



**Predicted risk calculated with the final model fitted on 2008 data**

|  |  |  |
| --- | --- | --- |
| **Deciles of Predicted Risk of Readmission** | **Observed Readmission Rate** | **Predicted Readmission Rate** |
| 1 | 7.8% | 8.9% |
| 2 | 10.3% | 10.6% |
| 3 | 12.0% | 11.9% |
| 4 | 13.9% | 13.2% |
| 5 | 14.6% | 14.5% |
| 6 | 16.8% | 15.9% |
| 7 | 19.1% | 17.6% |
| 8 | 20.7% | 19.9% |
| 9 | 23.5% | 23.2% |
| 10 | 31.5% | 32.2% |

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

Please see response in section 2b4.3 above.

**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*)

Please see response in section 2b4.3 above.

*Provide the statistical results from testing the approach to controlling for differences in patient characteristics (case mix) below*.  
***if stratified, skip to*** [***2b4.9***](#question2b49)

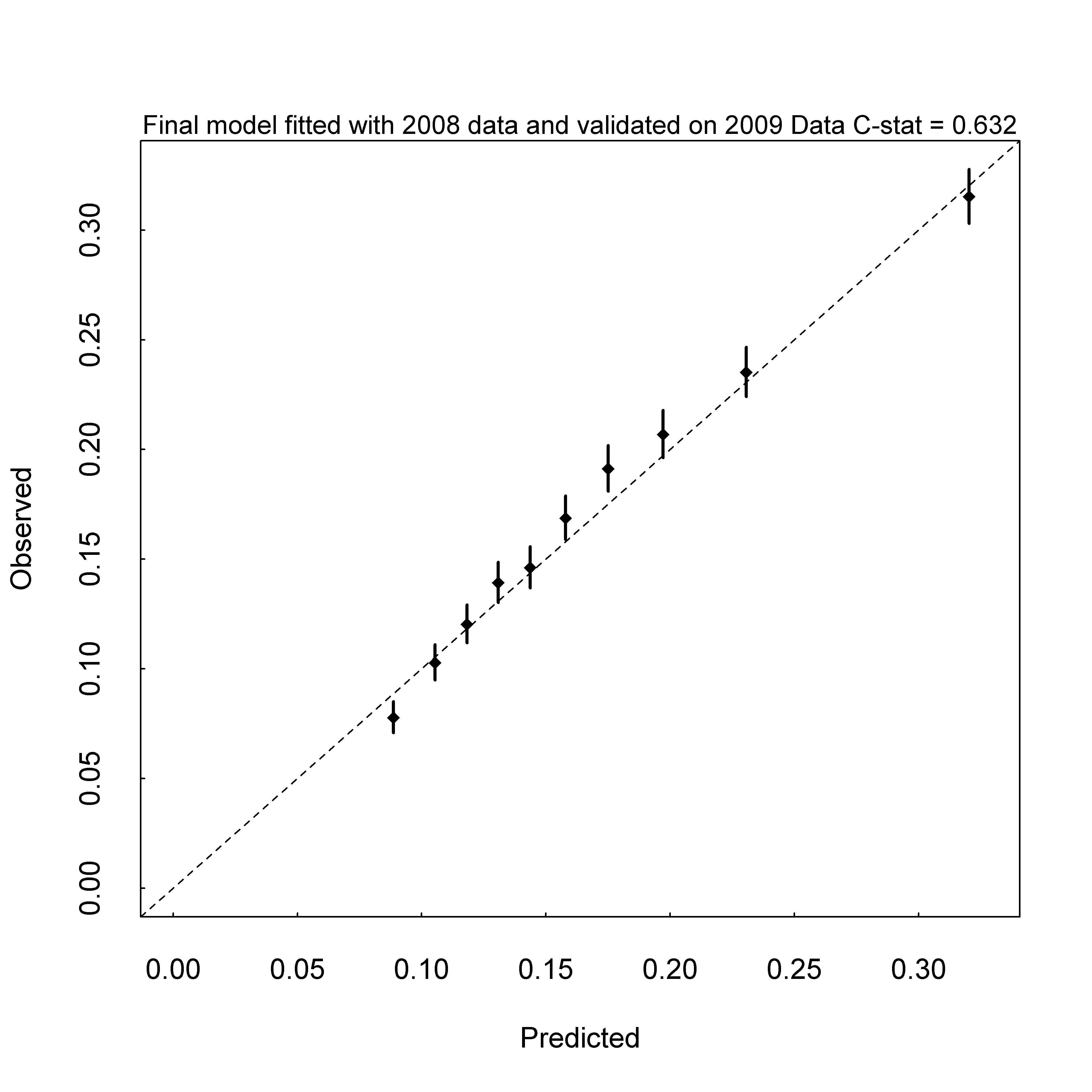
**2b4.6. Statistical Risk Model Discrimination Statistics** (*e.g., c-statistic, R-squared*)**:**

The C-index was 0.632. This level of discrimination is comparable to several other existing CMS readmission measures.

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

As shown in the figure and table below in 2b4.8, there was excellent agreement between observed and expected readmission rates.

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



**Predicted risk calculated with the final model fitted on 2008 data**

|  |  |  |
| --- | --- | --- |
| **Deciles of Predicted Risk of Readmission** | **Observed Readmission Rate** | **Predicted Readmission Rate** |
| 1 | 7.8% | 8.9% |
| 2 | 10.3% | 10.6% |
| 3 | 12.0% | 11.9% |
| 4 | 13.9% | 13.2% |
| 5 | 14.6% | 14.5% |
| 6 | 16.8% | 15.9% |
| 7 | 19.1% | 17.6% |
| 8 | 20.7% | 19.9% |
| 9 | 23.5% | 23.2% |
| 10 | 31.5% | 32.2% |

**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*)

Please see response in section 2b4.3 above.

\***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*)

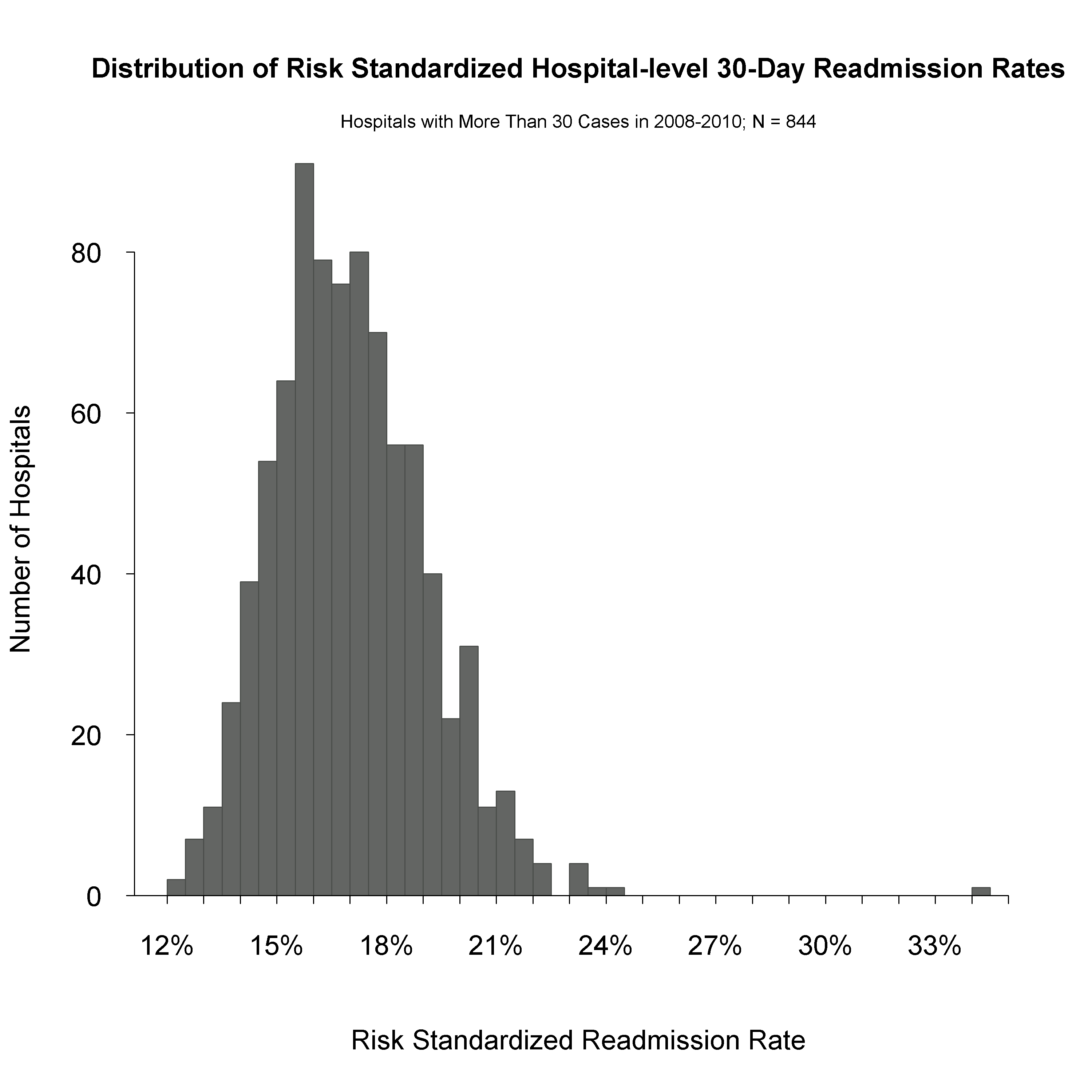
**2b5. IDENTIFICATION OF STATISTICALLY SIGNIFICANT & MEANINGFUL DIFFERENCES IN PERFORMANCE**

**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 proposed readmission measure was pilot tested by applying it to linked STS-CMS data from 2008-2010. The analysis population consisted of Medicare fee-for-service beneficiaries age 65 years or older at discharge who underwent isolated coronary artery bypass grafting (CABG) during 2008-2010 and were discharged alive. Candidate CABG admissions were identified by selecting Medicare Part A claims with an ICD-9-CM procedural code for CABG (36.1x) in any position. Records were retained for analysis if they met the following additional criteria: (1) linked to an STS record for isolated CABG (see below for record linkage criteria and definition of isolated CABG); (2) eligible for Medicare fee-for-service (FFS) A and B for at least two months after discharge or until month of death, whichever is first; (3) discharged from acute care setting within 1 year of index CABG admission; (4) did not leave against medical advice; (5) no logically inconsistent claims data (e.g. claims with overlapping admission and discharge dates); (6) is the first eligible operation per patient during the measurement period. After applying these inclusion/exclusion criteria, the analysis population included 162,503 index CABG admissions from 1010 CMS hospitals. Results are reported for the subset of 885 hospitals with at least 30 eligible cases during 2008-2010.

We calculated risk-standardized readmission rates (RSRR’s) for a subset of 885 hospitals having at least 30 eligible CABG cases during 2008-2010. We then summarized the distribution of RSRR’s across these hospitals.

**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*)



|  |  |
| --- | --- |
| **Distribution Statistic** | **Risk Standardized Readmission Rate** |
| **Number of Hospitals** | 844 |
| **Minimum** | 12.5% |
| **10%** | 14.5% |
| **25%** | 15.6% |
| **50%** | 16.9% |
| **75%** | 18.3% |
| **90%** | 19.8% |
| **Maximum** | 34.2% |
| **Mean** | 17.0% |

**2b5.3. What is your interpretation of the results in terms of demonstrating the ability to identify statistically significant and/or clinically/practically meaningful differences in performance across measured entities?** (i*.e., what do the results mean in terms of statistical and meaningful differences?*)

Number of Hospitals With 95% Interval Estimators Excluding the National Rate in 2008-2010 Data, Subset of Hospitals With At Least 30 Cases and Linkage Rate At Least 90%

|  |  |  |
| --- | --- | --- |
| **# better than  expected** | **# as  expected** | **# worse than  expected** |
| 25 | 794 | 27 |

For comparison, consider the current (February 3, 2014) Hospital Compare website display for readmission rates of acute MI patients discharged between July 1, 2009 and June 30, 2012. The overall U.S. national rate of readmission = 18.3%. Among 4,464 hospitals in the United States, 23 hospitals (0.5%) were Better than U.S. National Rate, 2,327 hospitals were No different than the U.S. National Rate, 29 (0.6%) were Worse than the U.S. National Rate, and 2,085 hospitals did not have enough cases to reliably tell how well they are performing.

**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 datasources/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