**National Quality Forum—Measure Testing (subcriteria 2a2, 2b1-2b6)**

**Measure Number** (*if previously endorsed*)**:** 0536

**Measure Title**: 30-day all-cause risk-standardized mortality rate following Percutaneous Coronary Intervention (PCI) for patients with ST segment elevation myocardial infarction (STEMI) or cardiogenic shock

**Date of Submission**: 11/8/2017

**Type of Measure:**

|  |  |
| --- | --- |
| Outcome (*including PRO-PM*) | Composite – ***STOP – use composite testing form*** |
| Intermediate Clinical Outcome | Cost/resource |
| Process *(including Appropriate Use)* | 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, 2b1, 2b2, and 2b4 must be completed.** * **For outcome and resource use measures**, section **2b3** also must be completed. * If specified for **multiple data sources/sets of specificaitons** (e.g., claims and EHRs), section **2b5** 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 (2b1-2b6) 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 25 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). * For information on the most updated guidance on how to address social risk factors variables and testing in this form refer to the release notes for version 7.1 of the Measure Testing Attachment. |

|  |
| --- |
| **Note:** The information provided in this form is intended to aid the Standing 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 **instrument-based measures** (including PRO-PMs) **and composite performance measures**, reliability should be demonstrated for the computed performance score.  **2b1.** **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 **instrument-based measures (including PRO-PMs) and composite performance measures**, validity should be demonstrated for the computed performance score.    **2b2.** **Exclusions** are supported by the clinical evidence and are of sufficient frequency to warrant inclusion in the specifications of the measure; [**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)  **2b3.** **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 (including clinical and social risk factors) that influence the measured outcome 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.   **2b4.** 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.  **2b5.** **If multiple data sources/methods are specified, there is demonstration they produce comparable results**.  **2b6.** 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. The degree of consensus and any areas of disagreement must be provided/discussed.  **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.** 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.17*)** | **Measure Tested with Data From:** |
| abstracted from paper record | abstracted from paper record |
| claims | claims |
| registry | registry |
| abstracted from electronic health record | abstracted from electronic health record |
| eMeasure (HQMF) implemented in EHRs | eMeasure (HQMF) implemented in EHRs |
| other: Source of vital status (e.g. National death index) | other: Source of vital status (e.g. National death index) |

**1.2. If an existing dataset was used, identify the specific dataset** (*the dataset used for testing must be consistent with the measure specifications for target population and healthcare entities being measured; e.g., Medicare Part A claims, Medicaid claims, other commercial insurance, nursing home MDS, home health OASIS, clinical registry*).

Medicare Part A claims, National Cardiovascular Data Registry (NCDR) CathPCI Registry,

Medicare Enrollment Database

We linked CathPCI Registry and Medicare data and identified in-hospital deaths using the discharge disposition indicator in the Standard Analytic File (SAF) and identified post-discharge deaths using the Enrollment Database (EDB)

**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.

**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.20*)** | **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*)

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:

Measure reliability and validity dataset

The measure reliability and validity dataset linked the CathPCI and Medicare Part A claims data from 2010-2011. It included 48,339 admissions to 1,182 hospitals with 24,170 admissions to 1,167 hospitals in one randomly selected sample and 24,169 admissions to 1,160 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 360 hospitals and the second hospital contained 360 hospitals. The linked dataset was used for:

- Data element reliability testing (Section 2a2)

- Measure score validity testing (Section 2b2)

- Measure exclusions testing (Section 2b3)

Data validity (Section 2b2)

We used admissions of patients discharged from January through December 2005.

Risk adjustment dataset (Section 2b4)

We use admissions with PCI in the merged data from 2006. The development sample consisted of 15,123 admissions at 602 hospitals in the STEMI or shock cohort.

**1.8** **What were the social risk factors that were available and analyzed**? For example, patient-reported data (e.g., income, education, language), proxy variables when social risk data are not collected from each patient (e.g. census tract), or patient community characteristics (e.g. percent vacant housing, crime rate) which do not have to be a proxy for patient-level data.

Social risk factors were not used in this risk model for the following reasons. First, as a detailed clinical registry used for quality assessment and improvement, there are not prospective interviews with patients to obtain patient-reported data. Second, the effect of social risk factors may be at either the patient- or the hospital-level. For example, patients with social risk factors (i.e., low income, lack of education, etc.) may have an increased risk of mortality because these patients may have an individual higher risk (patient-level effect) or because patients with social risk factors are more often admitted to hospitals with higher overall mortality rates (hospital-level effect). It is important to note, however, that even in the presence of a significant patient-level effect and absence of a significant hospital-level effect, the increased risk could be partly or entirely due to the quality of care patients receive in the hospital. For example, biased or differential care provided within a hospital to low-income patients as compared to high-income patients would exert its impact at the level of individual patients, and therefore be a patient-level effect. Third, while it may be true that worse social risk factors might be associated with more severe illness at the time of presentation, we had direct access to detailed clinical variables describing the severity of illness and feel that incorporating such factors (e.g. cardiogenic shock, ejection fraction, PCI status, cardiac arrest, highest risk legion, etc.) is a much more accurate means of stratifying risk. Accordingly, we feel that in this model of 30-day AMI mortality for STEMI/Shock patients, given the rich clinical data available through the NCDR CathPCI Registry and linkage to National Death Index data, that social risk factors, which are not readily available, would not likely contribute much improvement to this particular risk model.

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

**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

See Section 2b2 for validity testing of data elements

Measure Score Reliability

To assess reliability of the measure, we examined the extent to which assessments of a hospital using different but randomly selected subsets of patients in the same time period produced similar measures of hospital performance. That is, we took 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 calculated the agreement of the two resulting performance measures across hospitals.

For test-retest reliability of the measure in Medicare FFS patients aged 65 and older, we combined index admissions from two years (2010 and 2011) into a single dataset, randomly sampled half of patients within each hospital, calculated the measure for each hospital, and repeated the calculation using the second half. Thus, we measured each hospital 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 reliable. As a metric of agreement we calculated the intra-class correlation coefficient and assessed the values according to conventional standards.

Specifically, we used a combined 2010-2011 sample that had been linked with Medicare FFS claims data, and randomly split it into two approximately equal subsets of patients. We then calculated the RSMR for each hospital for each sample. The agreement of the two RSMRs was quantified for hospitals in each sample using the intra-class correlation. 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. Of note, because our final measure is derived using hierarchical logistic regression, a known property of hierarchical logistic regression models is that smaller 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%.

**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*)

Measure Score Reliability

We calculated the correlation of the RSMR from our final model in two different samples.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Table 1. Overall mortality rate (OMR) and risk-standardized mortality rate (RSMR) in the split samples; 2010-2011.** | | | | | | |
| **Description** | **First Half of the Data** | | | **Second Half of the Data** | | |
| **Volume** | **Weighted by Hospital Volume** | | **Volume** | **Weighted by Hospital Volume** | |
| **OMR** | **RSMR** | **OMR** | **RSMR** |
|  |  |  |  |  |  |  |
| N | 1,167 | 24,170 | 24,170 | 1,160 | 24,169 | 24,169 |
| Mean | 20.71 | 0.1230 | 0.1245 | 20.84 | 0.1209 | 0.1211 |
| Std Deviation | 17.48 | 0.0799 | 0.0249 | 17.19 | 0.0775 | 0.0152 |
|  |  |  |  |  |  |  |
| 100% Max | 136 | 1.0000 | 0.2274 | 141 | 1.0000 | 0.1754 |
| 99% | 90 | 0.3750 | 0.2041 | 84 | 0.3333 | 0.1663 |
| 95% | 52 | 0.2667 | 0.1700 | 55 | 0.2500 | 0.1498 |
| 90% | 42 | 0.2195 | 0.1575 | 43 | 0.2174 | 0.1413 |
| 75% Q3 | 28 | 0.1667 | 0.1374 | 28 | 0.1579 | 0.1299 |
| 50% Median | 17 | 0.1154 | 0.1216 | 17 | 0.1111 | 0.1189 |
| 25% Q1 | 9 | 0.0667 | 0.1067 | 9 | 0.0714 | 0.1103 |
| 10% | 3 | 0.0303 | 0.0952 | 3 | 0.0357 | 0.1037 |
| 5% | 2 | 0.0000 | 0.0905 | 2 | 0.0000 | 0.0993 |
| 1% | 1 | 0.0000 | 0.0799 | 1 | 0.0000 | 0.0940 |
| 0% Min | 1 | 0.0000 | 0.0703 | 1 | 0.0000 | 0.0888 |

**Figure 1. Correlation between Hospital Risk-Standardized Mortality Rates in Split Samples; 2010-2011.**

****

**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 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. Accordingly, our approach to assessing reliability is to consider the extent to which assessments of a hospital using different but randomly selected subsets of patients in the same time period produce similar measures of hospital performance. The agreement between the two RSMRs (0.122), which according to conventional interpretation is “slight,” likely reflects the relatively low number of cases included in the cohort as outlined above (Landis JR et al. 2013). Nevertheless, the reliability of the measure should be assessed using larger split samples when available. Based on our experience with similar measures using split samples, using 4 years (and volume equivalent to 2 years) would result in higher intra-class correlation coefficient.

References

Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. Mar 1977;33(1):159-174.

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

**2b1. VALIDITY TESTING**

**2b1.1. 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*) **NOTE**: Empirical validity testing is expected at time of maintenance review; if not possible, justification is required.

**2b1.2. 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)*

Data Element Validity

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 Data Quality Program 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. 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.

10th revision of the International Statistical Classification of Diseases and Related Health Problems (ICD-10) Code Selection

In 2012, we used the General Equivalence Mapping (GEM) crosswalk between ICD-9-CM and ICD-10-CM/PCS to create specifications for the measure in ICD-10-CM/PCS. Our process for mapping procedural codes in the measures to ICD-10-CM consisted of a detailed clinical review, including manual review of related ICD-10-CM codes to determine that all appropriate codes are included, rather than relying exclusively on the GEM. To conduct the crosswalk, we created a database to effectively use the mapping tables provided by CMS. We then compiled a list of ICD-9-CM codes that define PCI during hospitalization. Measure developers used these ICD-9-CM codes to build queries to extract the GEM results from the mapping table in the database. We then applied those ICD-10-CM codes to the ICD-10-CM to ICD-9-CM mapping table to see if the reverse query produced ICD-9-CM codes that were not in the original measure specifications.

Our clinicians reviewed these results in detail and determined that many ICD-10-CM codes that should be included in our cohort were not being captured by the GEMs. We confirmed this by consulting the ICD-10-CM draft procedural codebook and identifying the ICD-10-CM codes that our clinicians felt should be included in our cohort. The GEMs identified 16 ICD-10-CM codes for our PCI mortality cohort, while clinician review of the ICD-10-CM draft codebook resulted in 48 ICD-10-CM codes.

Further details also are located in the attached Appendix.

**2b1.3. What were the statistical results from validity testing**? (*e.g., correlation; t-test*)

Data Element Validity

In the audit that assessed cases submitted in 2005, the median agreement between submitted and audited values was 92%. There was consistency across sites, with agreement in the lowest and highest deciles of hospitals ranging from 90% to 95%.

**2b1.4. 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?*)

Data Element Validity

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.

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

**2b2. EXCLUSIONS ANALYSIS**

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

**2b2.1. Describe the method of testing exclusions and what it tests** (*describe the steps―do not just name a method; what was tested, e.g., whether exclusions affect overall performance scores; what statistical analysis was used*)

Exclusions were those determined by expert input to be clinically relevant, required in order to assess the outcome, or needed for calculation of the measure. To ascertain the impact of the exclusions on the cohort, we examined proportions of the total cohort excluded for each exclusion criterion

**2b2.2. 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*)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Table 2. Exclusions from the target population for the 2010, 2011, and the combined 2010-2011 dataset.** | | | | | |
| **Exclusions** | **2010** | | **2011** | | **2010-2011** |
| **Patient Stay** | **Hospitals** | **Patient Stay** | **Hospitals** | **Patient Stay** |
| **#**  **(%)** | **#**  **(%)** | **#**  **(%)** | **#**  **(%)** | **#**  **(%)** |
| **Initial Sample** | **199,853** | **1,095** | **195,812** | **1,185** | **395,665** |
| Not Medicare patient on admission | 43, 669 (21.85) | 3  (0.27) | 44,840  (22.90) | 1  (0.08) | 88509  (22.37) |
| **Remaining** | **156184** | **1,092** | **150,972** | **1,184** | **307,156** |
| Not the first claim in the same claim bundle\* | 3  (0.00) | 0  (0.00) | 8 (0.01) | 0  (0.00) | 11  (0.00) |
| **Remaining** | **156181** | **1,092** | **150,964** | **1,184** | **307,145** |
| Get the procedure more than 10 days after admission | 1,074 (0.69) | 0  (0.00) | 1,212  (0.80) | 0  (0.00) | 2286  (0.74) |
| **Remaining** | **155,107** | **1,092** | **149,752** | **1,184** | **304,859** |
| Transferred in (PCI to PCI) | 186 (0.12) | 0  (0.00) | 204  (0.14) | (0.00) | 390  (0.13) |
| **Remaining** | **154,921** | **1,092** | **149,548** | **1,184** | **304,469** |
| Unknown death | 0  (0.00) | 0  (0.00) | 0  (0.00) | 0  (0.00) | 0  (0.00) |
| **Remaining** | **154,921** | **1,092** | **149,548** | **1,184** | **304,469** |
| Duplicate death | 65  (0.04) | 0  (0.00) | 77  (0.05) | 0  (0.00) | 142  (0.05) |
| **Remaining** | **154,856** | **1,092** | **149,471** | **1,184** | **304,327** |
| AMA | 215 (0.14) | 0  (0.00) | 212  (0.14) | 0  (0.00) | 427  (0.14) |
| **Remaining** | **154,641** | **1,092** | **149,259** | **1,184** | **303,900** |
| Not with STEMI/Shock | 130,942 (84.67) | 20  (1.83) | 124619  (83.49) | 28  (2.36) | 255561  (84.09) |
| **Study Sample** | **23,699** | **1,072** | **24,640** | **1,156** | **48,339** |
| Death within 30-days from procedure | 2,804 (11.83) |  | 3090  (12.54) |  | 5894  (12.19) |
| In-Hospital death | 2,241 (9.5) |  | 2439  (9.9) |  | 4680  (9.68) |
| \* 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. | | | | | |

**2b2.3. 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 decision to exclude patients discharged AMA is based on clinical judgment to make the measure fair and is unlikely to distort the results given the very low frequency. Excluding patients transferring into a hospital does not actually exclude acute episodes from the measure, but considers the hospital that initially admits the patient as the one accountable for the outcome, avoiding double counting and clarifying accountability. The exclusion of unreliable data is necessary for valid calculation of the measure. Excluding PCIs that follow a prior PCI in the same admission or during a transfer-in is applied in order to avoid assigning the death to two separate admissions. The decision to exclude subsequent PCIs within 30 days of death is necessary to avoid attributing the same death to more than one PCI. Lastly, patients who get the procedure more than 10 days after admission have a PCI after many days of hospitalization are rare and represent a distinct population that likely has risk factors related to the hospitalization and not well quantified in the registry.

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

**2b3. 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)***.***

**2b3.1. What method of controlling for differences in case mix is used?**

**No risk adjustment or stratification**

**Statistical risk model with** 13 **risk factors**

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

**Other,** Click here to enter description

**2b3.1.1 If using a statistical risk model, provide detailed risk model specifications, including the risk model method, risk factors, coefficients, equations, codes with descriptors, and definitions.**

**2b3.2. If an outcome or resource use component 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**

**2b3.3a. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors (clinical factors or social risk 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*) **Also discuss any “ordering” of risk factor inclusion**; for example, are social risk factors added after all clinical factors?

The goal of risk adjustment is to account for different patient demographic and clinical characteristics at the time of admission (hospital case mix), enabling interpretation of any identified differences in quality. Conditions that may represent adverse outcomes due to care received during the index hospital stay are not included in the risk-adjustment model. We sought to develop a model that included key variables that were clinically relevant and based on strong association with 30-day mortality.

**2b3.3b. How was the conceptual model of how social risk impacts this outcome developed? Please check all that apply:**

**Published literature**

**Internal data analysis**

**Other (please describe)**

**2b3.4a. What were the statistical results of the analyses used to select risk factors?**

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, or discharged to, a skilled nursing facility [SNF]). Variables were also considered ineligible if they were particularly vulnerable to gaming or were deemed to lack clinical relevance. Based on careful review by a team of clinicians and further informed by a review of the literature, a total of 26 variables were determined to be appropriate for consideration as candidate variables. Our set of candidate variables included two “demographic” variables (age and gender), 15 “history and risk factor” variables, four “cardiac status” variables, one “cath lab visit” variable and four “PCI procedure” variables. The final risk-adjustment model for the STEMI or shock cohort included 13 variables:

1) Age

2) Body mass index (BMI)

3) Cerebrovascular disease

4) Chronic lung disease

5) Glomerular filtration rate (GFR)

6) Previous PCI

7) Congestive heart failure (CHF) status

8) Cardiogenic shock

9) Symptoms present on admission

10) Ejection fraction percentage

11) PCI status

12) Highest risk lesion – segment category

13) Highest risk lesion – SCAI lesion class

**2b3.4b. Describe the analyses and interpretation resulting in the decision to select social risk factors** *(e.g. prevalence of the factor across measured entities, empirical association with the outcome, contribution of unique variation in the outcome, assessment of between-unit effects and within-unit effects.)* **Also describe the impact of adjusting for social risk (or not) on providers at high or low extremes of risk.**

**N/A**

**2b3.5. 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*)

Several variables required particular consideration. First, in the current version of the CathPCI registry, participants are instructed to use New York Heart Association (NYHA) classification to capture symptom severity for both heart failure and angina. Accordingly, the resulting variable is a hybrid which may dilute the prognostic importance usually associated with NYHA class. Second, variables such as PCI status and cardiogenic shock impart important prognostic information but are vulnerable to systematic misclassification. This is relevant to efforts to publicly report 30-day PCI mortality in that several key variables (e.g., cardiogenic shock and PCI status) may be consistently coded differently across sites. For example, although the CathPCI data dictionary provides detailed definitions of PCI status (http://www.ncdr.com/WebNCDR/ELEMENTS.ASPX), sites may differ in their interpretation of these definitions such that a patient considered an emergent PCI at hospital A may be considered an urgent PCI at hospital B. If differences in coding occur with sufficient frequency, the risk-standardized mortality rate for hospital A might appear lower than hospital B, even if their case mixes and outcomes were otherwise identical.

To examine this issue, we compared the frequency of different PCI status categories at hospitals with risk adjusted mortality rates that were above and below the median using the STEMI or shock cohort. We found that rates of cardiogenic shock were comparable, but that hospitals with below average risk-standardized mortality had modestly higher rates of emergency and salvage PCI (76.7% and 1.4%), compared with hospitals with above average risk-standardized mortality (72.3% and 1.2%). We cannot determine whether these differences accurately reflect differences in case mix or are due to systematic differences in coding. Nevertheless, these results highlight the need to further ensure data accuracy.

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 significant numbers of missing values: body mass index (BMI), glomerular filtration rate (GFR), and left ventricular ejection fraction (LVEF). For BMI, we stratified by gender and imputed the missing values to the median of the corresponding groups. For GFR, we stratified patients into five categories: <30, 31-60, 61-90, >90, and missing. For LVEF, we stratified patients into four categories- <30%, 31-45%, >45%, and missing.

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.

*Provide the statistical results from testing the approach to controlling for differences in patient characteristics (case mix) below*.  
***If stratified, skip to*** [***2b3.9***](#question2b49)

**2b3.6. Statistical Risk Model Discrimination Statistics** (*e.g., c-statistic, R-squared*)**:**We computed 6 summary statistics for assessing model performance: over-fitting indices, percentage of variation explained by the risk factors, predictive ability, area under the receiver operating characteristic (ROC) curve, distribution of residuals, and model chi-square.

The development model has excellent discrimination, calibration, and fit. The patient-level mortality rate ranges from 1.4% in the lowest predicted decile to 40.3% in the highest predicted decile, a range of 38.9%. The area under the ROC curve is 0.825.

The discrimination and the explained variation of the model at the patient-level are consistent with those of published PCI in-hospital mortality models (Yale-CORE 2008). The ROC is modestly lower than that of previously published models due to several factors. First, we stratified the entire population of PCI patients into two populations based on the presence or absence of two prognostically important variables: STEMI and cardiogenic shock. Second, we excluded covariates such as potential complications, certain patient demographics (e.g., race), and patients‟ admission path (e.g., outpatient, emergency department, transfers-in from other facilities (non-acute care or acute care). These characteristics may be associated with mortality and thus could increase the model performance to predict patient mortality. However, these variables may be related to quality or supply factors that should not be included in an adjustment that seeks to control for patient clinical characteristics. Thus, the choice was to focus on adjustment for clinical differences in the populations among hospitals.

**2b3.7. Statistical Risk Model Calibration Statistics** (*e.g., Hosmer-Lemeshow statistic*):

|  |  |  |
| --- | --- | --- |
| **Table 3. Model Performance: Calibration Results Based on the Logistic Regression Model** | | |
| **Indices** | **2011 Sample** | **2010 Sample** |
| Number of Admissions | 24,640 | 23,699 |
| Calibration |  |  |
| γ0, γ1 | - | -0.088, 0.997 |
| ROC | 0.827 | 0.831 |
| Residuals Lack of Fit (Pearson Residual Fall %) |  |  |
| <-2 | 0.175 | 0.186 |
| [-2, 0) | 87.285 | 87.983 |
| [0, 2) | 7.171 | 6.675 |
| [2+ | 5.369 | 5.156 |

**2b3.8. Statistical Risk Model Calibration – Risk decile plots or calibration curves**:

|  |  |  |  |
| --- | --- | --- | --- |
| **Table 4. RSMR Model Performance for STEMI or Shock Cohort** | | |  |
| **Indices** | **Development Sample**  **(2010)** | **Validation Sample**  **(2011)** | **Merged Sample**  **(2010-2011)** |
| Number of hospitals | 1,072 | 1,156 | 1,182 |
| Number of admissions | 23,699 | 24,640 | 48,339 |
| RSMR |  |  |  |
| 100% Max | 0.2077 | 0.2275 | 0.1983 |
| 99% | 0.1747 | 0.1906 | 0.1813 |
| 95% | 0.1578 | 0.1661 | 0.1616 |
| 90% | 0.1452 | 0.1542 | 0.1496 |
| 75% | 0.1297 | 0.1378 | 0.1331 |
| 50% Median | 0.1159 | 0.1249 | 0.1201 |
| 25% | 0.1060 | 0.1113 | 0.1094 |
| 10% | 0.0977 | 0.1007 | 0.0992 |
| 5% | 0.0936 | 0.0953 | 0.0935 |
| 1% | 0.0834 | 0.0841 | 0.0848 |
| 0% Min | 0.0751 | 0.0741 | 0.0778 |

**2b3.9. Results of Risk Stratification Analysis**:

N/A

**2b3.10. What is your interpretation of the results in terms of demonstrating adequacy of controlling for differences in patient characteristics (case mix)?** (i*.e., what do the results mean and what are the norms for the test conducted*)

The C-statistic of 0.825 indicates excellent model discrimination. The calibration value of close to 0 at one end and close to 1 to the other end indicates good calibration of the model. The risk decile plot shows excellent discrimination of the model and good predictive ability.

**2b3.11.** **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**

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

**2b4. IDENTIFICATION OF STATISTICALLY SIGNIFICANT & MEANINGFUL DIFFERENCES IN PERFORMANCE**

**2b4.1. 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). However, the decision to publicly report this PCI mortality measure and the approach to discriminating performance has not been determined.

We assessed variation in RSMRs among hospitals by examining the distribution of the hospital RSMRs and plotting the histogram of the hospital RSMRs.

**2b4.2. 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*)

In the 2010-2011 sample, the mean hospital RSMR for the STEMI or shock cohort was 12.3%%, with a range of 7.8% to 19.8%. The interquartile range was 10.9% to 13.3%.

Figure 2. Distribution of risk-standardized mortality rates (RSMRs); 2010-2011 combined sample.



**2b4.3. 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 meaningful differences across hospitals in the 30-day risk-standardized mortality after PCI in the STEMI or shock cohort.

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

**2b5. COMPARABILITY OF PERFORMANCE SCORES WHEN MORE THAN ONE SET OF SPECIFICATIONS**

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

**Note***: This item is directed to measures that are risk-adjusted (with or without social risk factors)* ***OR*** *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).* ***Comparability is not required when comparing performance scores with and without social risk factors in the risk adjustment model. However, if comparability is not demonstrated for measures with more than one set of specifications/instructions, the different specifications (e.g., for medical records vs. claims) should be submitted as separate measures.***

N/A

**2b5.1. Describe the method of testing conducted to compare 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

**2b5.2. 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

**2b5.3. What is your interpretation of the results in terms of the differences in 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

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

**2b6. MISSING DATA ANALYSIS AND MINIMIZING BIAS**

**2b6.1. 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.

**2b6.2. 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 significant numbers of missing values: body mass index (BMI), glomerular filtration rate (GFR), and left ventricular ejection fraction (LVEF). The frequency of missingness by hospital appeared to be evenly distributed across hospitals. Model performance and estimates of hospital RSMR were not significantly different when repeated excluding cases with missing data. The fact that the data was missing did not appear to be at random in that patients with missing data regarding GFR, and LVEF were at higher risk of death than those without missing data. Accordingly we created a dummy variable to capture that information.

For categorical variables with missing values, the value from the reference group was added. For BMI, we stratified by gender and imputed the missing values to the median of the corresponding groups. For GFR, we stratified patients into five categories: <30, 31-60, 61-90, >90, and missing. For LVEF, we stratified patients into four categories- <30%, 31-45%, >45%, and missing.

**2b6.3. 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.