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

**Measure Number** (*if previously endorsed*)**:** 0133

**Measure Title**: In-Hospital Risk Adjusted Rate of Mortality for Patients Undergoing PCI

**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: 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*).

We propose to use a clinical registry, the National Cardiovascular Data Registry (NCDR) for CathPCI Registry. This is a national quality improvement registry that is currently participated in >1,300 US hospitals. Some states and healthcare systems mandate participation. Rigorous quality standards are applied to the data and both quarterly and ad hoc performance reports are generated for participating centers to track and improve their performance.

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

Since this model has already been approved by NQF as a performance measure, we have performed additional testing, in new data, to establish its continued value and accuracy as a performance measure.

We have chosen to use different datasets to provide support for different aspects of the proposed measure.

1. Audit data: 01/2009-12/ 2009 has been previously used to support the inter-rater reliability of the application. It was established that there is high inter-rater reliability, as compared with independent chart audits and found >90% accuracy for most variables. Please see prior submission for these data.

2. Creation of the Mortality model was performed on all national NCDR data from 07/2009–06/2011 and has been used to provide a description and initial performance characteristics of the model.

3. A separate cohort of the NCDR CathPCI registry was used to validate the model, which included all data collected during the 2012 calendar year (01/2012-12/2012). These data were also used to provide test-retest reliability of the data elements for the risk model and further validation of the relationship between the predictor variables and mortality, including additional data supporting the discrimination and calibration of the model.

**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*)

*Creation of the Mortality Derivation and Validation model:*

1,253 hospitals were included. See additional information under section 1.6.

**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*)   
For the updated derivation and validation of the mortality risk model, 1,208,137 patients undergoing PCI between 7/2009-06/2011 at 1,253 hospitals were included; 60% in the derivation cohort and a random 40% in the validation cohort. In-hospital mortality was 1.4%, ranging from 0.2% among elective cases (45.1% of total cases) to 65.9% among patients with shock and recent cardiac arrest (0.2% of total cases) . A summary of these patients’ clinical characteristics and the hospital characteristics are provided under Table 1 and 2.

Table 1. Derivation and Validation Characteristics

|  |  |  |
| --- | --- | --- |
|  | **Development (N=724,883)** | **Validation (N=483,254)** |
| Overall population | 1.38 | 1.40 |
| MI status |  |  |
| STEMI | 5.22 | 5.33 |
| No STEMI | 0.65 | 0.65 |
| Gender |  |  |
| Men | 1.22 | 1.23 |
| Women | 1.72 | 1.74 |
| Age group |  |  |
| Age > 70 yrs | 2.23 | 2.25 |
| Age ≤ 70 yrs | 0.96 | 0.98 |
| Diabetes status |  |  |
| Diabetes mellitus | 1.51 | 1.50 |
| No diabetes mellitus | 1.31 | 1.34 |
| Cardiac arrest | 24.32 | 25.07 |
| Cardiogenic shock and PCI status |  |  |
| Sustained shock and salvage | 63.99 | 68.85 |
| Sustained shock or salvage | 33.45 | 34.33 |
| Transient shock but not salvage | 15.26 | 14.85 |
| Emergency PCI without shock/salvage | 2.26 | 2.29 |
| Urgent PCI without shock/salvage | 0.63 | 0.63 |
| Elective PCI without shock/salvage | 0.18 | 0.18 |
| STEMI = ST-segment elevation myocardial infarction; All other abbreviations can be found in Table 1. | | |

Table 2. Hospital Characteristics

| 2012 Data | |
| --- | --- |
|  | Total |
| n = 1367 |
| Participant Classification      FREE STANDING CATH LAB      FREE STANDING CATH LAB/CLINIC      HEALTH SYSTEM/NETWORK      HOSPITAL      HOSPITAL/HEALTH NETWORK      OTHER      PRIVATE CV PRACTICE      Missing | 1 (0.1%) 3 (0.2%) 60 (4.4%) 1203 (88.1%) 95 (7.0%) 3 (0.2%) 1 (0.1%) 1 |
| Hospital Location      RURAL      SUBURBAN      URBAN      Missing | 249 (18.2%) 492 (36.0%) 625 (45.8%) 1 |
| Participant Type      GOVERNMENT      PRIVATE/COMMUNITY      UNIVERSITY      Missing | 21 (1.5%) 1232 (90.2%) 113 (8.3%) 1 |
| Teaching Hospital      Missing (.) | 524 (38.4%) 1 |
| Public Hospital      Missing (.) | 530 (38.8%) 1 |
| Volume (Med (IQR)) | 367 (188, 643) |
| Census Region      MIDWEST REGION      NORTHEAST REGION      SOUTH REGION      WEST REGION      Missing | 395 (28.9%) 182 (13.3%) 521 (38.2%) 267 (19.6%) 2 |

For the additional testing of predictive validity, calibration and test-retest reliability, we used 634,084 patients undergoing PCI between 1/2012-12/2012, of whom 10,212 (1.6%) had a mortality event. A summary of these patients’ clinical characteristics (focusing upon those that are predictor variables in the final, full model) are provided under Table 3.

Table 3. Predicted Probability of Mortality

|  | | | | |
| --- | --- | --- | --- | --- |
|  | Total | Observed Mortality | | P-Value |
| n = 634084 | Yes n = 10212 | No n = 623872 |
| ***Mortality*** |  |  |  |  |
| Predicted Probability of Death | 0.01583 ± 0.06557 | 0.26718 ± 0.27434 | 0.01171 ± 0.04569 | < 0.001 |
| ***Mortality Variables*** |  |  |  |  |
| STEMI | 111775 (17.6%) | 6275 (61.4%) | 105500 (16.9%) | < 0.001 |
| Age | 64.8 ± 12.1 | 70.4 ± 12.8 | 64.7 ± 12.0 | < 0.001 |
| Body Mass Index | 30.0 ± 6.4 | 28.5 ± 7.0 | 30.0 ± 6.4 | < 0.001 |
| CVD | 79750 (12.6%) | 1771 (17.3%) | 77979 (12.5%) | < 0.001 |
| PVD | 79224 (12.5%) | 1800 (17.6%) | 77424 (12.4%) | < 0.001 |
| Prior PCI | 258993 (40.8%) | 2665 (26.1%) | 256328 (41.1%) | < 0.001 |
| Left Ventricular Ejection Fraction | 52.6 ± 10.1 | 43.2 ± 13.6 | 52.8 ± 10.0 | < 0.001 |
| GFR | 71.4 ± 18.0 | 55.9 ± 19.9 | 71.7 ± 17.9 | < 0.001 |
| Non-insulin Diabetes vs. No Diabetes | 145444 (22.9%) | 2256 (22.1%) | 143188 (23.0%) | 0.040 |
| Insulin Diabetes vs. No Diabetes | 91052 (14.4%) | 1811 (17.7%) | 89241 (14.3%) | < 0.001 |
| HF NYHA Class I/II/III w/in 2 Wks vs. No HF | 49822 (7.9%) | 1250 (12.2%) | 48572 (7.8%) | < 0.001 |
| HF NYHA Class IV w/in 2 Wks vs. No HF | 16595 (2.6%) | 1889 (18.5%) | 14706 (2.4%) | < 0.001 |
| Cardiac arrest w/in 24 hrs | 13637 (2.2%) | 3559 (34.9%) | 10078 (1.6%) | < 0.001 |
| Salvage Status and Carshock w/in 24hrs and at start of PCI vs. Elective Status and No Carshock | 1563 (0.2%) | 1072 (10.5%) | 491 (0.1%) | < 0.001 |
| Salvage Status or Carshock w/in 24hrs and at start of PCI (not both) vs. Elective Status and No Carshock | 9115 (1.4%) | 3159 (30.9%) | 5956 (1.0%) | < 0.001 |
| Carshock w/in 24hrs or at start of PCI (not both) vs. Elective Status and No Carshock | 8460 (1.3%) | 1295 (12.7%) | 7165 (1.1%) | < 0.001 |
| Emergent Status and No Carshock vs. Elective Status and No Carshock | 107052 (16.9%) | 2440 (23.9%) | 104612 (16.8%) | < 0.001 |
| Urgent Status and No Carshock vs. Elective Status and No Carshock | 254411 (40.1%) | 1757 (17.2%) | 252654 (40.5%) | < 0.001 |
| pLAD vs. Other | 107452 (16.9%) | 2732 (26.8%) | 104720 (16.8%) | < 0.001 |
| Left Main vs. Other | 14756 (2.3%) | 831 (8.1%) | 13925 (2.2%) | < 0.001 |
| In-stent Thrombosis on some lesion previously treated w/in 1 month | 2044 (0.3%) | 152 (1.5%) | 1892 (0.3%) | < 0.001 |
| Number of Diseased Vessels (2,3) vs. (0,1) | 260307 (41.1%) | 6245 (61.2%) | 254062 (40.7%) | < 0.001 |
| Chronic Total Occlusion | 19138 (3.0%) | 597 (5.8%) | 18541 (3.0%) | < 0.001 |
| ***History*** |  |  |  |  |
| Intra-Aortic Balloon Pump           Missing (.) | 15357 (2.4%) 169 | 3864 (37.8%) 1 | 11493 (1.8%) 168 | < 0.001 |
| Prior MI           Missing (.) | 193054 (30.5%) 173 | 2808 (27.6%) 21 | 190246 (30.5%) 152 | < 0.001 |
| Prior PCI           Missing (.) | 258993 (40.9%) 150 | 2665 (26.1%) 13 | 256328 (41.1%) 137 | < 0.001 |
| Currently on Dialysis           Missing (.) | 15882 (2.5%) 605 | 665 (6.5%) 21 | 15217 (2.4%) 584 | < 0.001 |
| Chronic Lung Disease           Missing (.) | 97244 (15.3%) 322 | 2202 (21.6%) 23 | 95042 (15.2%) 299 | < 0.001 |
| Diabetes Mellitus           Missing (.) | 236496 (37.3%) 297 | 4067 (39.9%) 16 | 232429 (37.3%) 281 | < 0.001 |
| ***Cath Lab Visit*** |  |  |  |  |
| PCI Indication           Immediate PCI for STEMI           PCI for STEMI (Unstable, >12 hrs from Sx onset)           PCI for STEMI (Stable, >12 hrs from Sx onset)           PCI for STEMI (Stable after successful full-dose Thrombolysis)           Rescue PCI for STEMI (after failed full-dose lytics)           PCI for high risk Non-STEMI or unstable angina           Staged PCI           Other           Missing (.) | 97691 (15.4%) 5944 (0.9%) 2546 (0.4%) 2221 (0.4%) 3364 (0.5%) 332909 (52.5%) 34929 (5.5%) 154318 (24.3%) 162 | 5516 (54.0%) 498 (4.9%) 72 (0.7%) 15 (0.1%) 172 (1.7%) 3231 (31.6%) 98 (1.0%) 607 (5.9%) 3 | 92175 (14.8%) 5446 (0.9%) 2474 (0.4%) 2206 (0.4%) 3192 (0.5%) 329678 (52.9%) 34831 (5.6%) 153711 (24.6%) 159 | < 0.001 |
| CAD Presentation           No symptom, no angina           Symptom unlikely to be ischemic           Stable angina           Unstable angina           Non-STEMI           ST-Elevation MI (STEMI) or equivalent           Missing (.) | 35865 (5.7%) 14307 (2.3%) 89810 (14.2%) 249827 (39.4%) 134840 (21.3%) 109286 (17.2%) 149 | 328 (3.2%) 119 (1.2%) 137 (1.3%) 934 (9.1%) 2539 (24.9%) 6152 (60.3%) 3 | 35537 (5.7%) 14188 (2.3%) 89673 (14.4%) 248893 (39.9%) 132301 (21.2%) 103134 (16.5%) 146 | < 0.001 |
| Heart Failure w/in 2 Weeks           Missing (.) | 66417 (10.5%) 275 | 3139 (30.8%) 10 | 63278 (10.1%) 265 | < 0.001 |
| Cardiomyopathy or Left Ventricular Systolic Dysfunction           Missing (.) | 68481 (10.8%) 156 | 2274 (22.3%) 2 | 66207 (10.6%) 154 | < 0.001 |
| Pre-operative Evaluation Before Non-Cardiac Surgery           Missing (.) | 11823 (1.9%) 223 | 101 (1.0%) 2 | 11722 (1.9%) 221 | < 0.001 |
| Cardiogenic Shock w/in 24 Hours           Missing (.) | 13197 (2.1%) 109 | 4528 (44.3%) | 8669 (1.4%) 109 | < 0.001 |
| Cardiac Arrest w/in 24 Hours           Missing (.) | 13637 (2.2%) 178 | 3559 (34.9%) 3 | 10078 (1.6%) 175 | < 0.001 |
| Pre-PCI Left Ventricular Ejection Fraction           Missing | 52.4 ± 12.5 188804 | 37.8 ± 16.1 5396 | 52.5 ± 12.4 183408 | < 0.001 |
| ***Outcomes*** |  |  |  |  |
| Discharge Status           Alive           Deceased | 623872 (98.4%) 10212 (1.6%) | 0 (0.0%) 10212 (100.0%) | 623872 (100.0%) 0 (0.0%) | < 0.001 |
| Continuous variables compared using Student's T-test. Categorical variables compared using chi-square or Fisher's exact test. | | | | |

**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**.

We used the same data described above for all aspects of this supplement, except for the test-retest reliability of the data elements, where we restricted the sample to those with 2 procedures in 2012.

**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 clincial registry used for quality assessment and improvement, there are not prospective interviews with patients to obtain patient-reported data. Second, while proxy variables could be considered, these were not felt to be relevant to an inpatient mortality model, in contrast to a longer-term outcome model where difficulties with access to care, affording medications or cardiac rehabilitation would be more important. Moreover, 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, cardiac arrest, etc.) is a much more accurate means of stratifying risk. Accordingly, we feel that in this model of in-hospital mortality, given the rich clinical data available through the NCDR CathPCI registry, 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*)

Measure Score:

ACCF performed the signal-to-noise analysis on the same cohort of individuals as noted under Section 1.3. (testing method 3). Only hospitals with a minimum of 10 eligible patients were included in the analysis to prevent undetected bias introduced by the inclusion of hospitals with a small sample size.

Data Element:

ACCF evaluated the test-retest reliability by reviewing CathPCI patients who were readmitted or had a repeat procedure in 2012. This approach enabled us to examine 2 independent abstractions of data for the same patient. For certain characteristics that would not change (e.g. gender), we would expect near perfect reproducibility. For other characteristics (e.g. diabetes) we would expect that any patient diagnosed with diabetes on the first visit should also have diabetes recorded on the second visit. It is, however, clinically plausible that someone could be diagnosed with diabetes between their first and second visit, so the emergence of diabetes on the second visit is not necessarily an ‘error’ and no interpretation is made for these scenarios.

Data Element:

The NCDR Data Quality Program ensures that data submitted to the NCDR are complete validly collected. The NCDR Data Quality Program consists of 3 main components: data completeness, consistency, and accuracy. Completeness focuses on the proportion of missing data within fields, whereas consistency determines the extent to which logically related fields contain values consistent with other fields. Accuracy characterizes the agreement between registry data and the contents of original charts from the hospitals submitting data. Before entering the Enterprise Data Warehouse (EDW), all submissions are scored for file integrity and data completeness, receiving 1 of 3 scores that are transmitted back to facilities using a color coding scheme. A “red light” means that a submission has failed because of file integrity problems such as excessive missing data and internally inconsistent data. Such data are not processed or loaded into the EDW. A “yellow light” status means that a submission has passed the integrity checks but failed in completeness according to predetermined thresholds. Such data are processed and loaded into the EDW but are not included in any registry aggregate computations until corrected. Facilities are notified about data submission problems and provided an opportunity to resubmit data. Finally, a “green light” means that a submission has passed all integrity and quality checks. Such submissions are loaded to the EDW. After passing the DQR, data are loaded into a common EDW that houses data from all registries and included for all registry aggregate computations. In a secondary transaction process, data are loaded into registry-specific, dimensionally modeled data marts. A summary of the Program is noted under Table 4.

Table 4. Data Quality Program Overview

|  |  |
| --- | --- |
| **Methodology** | * Nationwide program (i.e., all submitting participants in the   United States)   * Review of data submitted the previous year * Review of a subset of data elements that can rotate each year * Remote review of data combined with couple of onsite visit * Onsite visits are targeted based on the Data Outlier Program * Random selection of sites and records * Blinded data abstraction from medical charts * Inter-rater Reliability Assessment conducted to validate the audit   findings   * Adjudication step for participant to refute audit findings |
| **Scope** | * Review of hospital’s medical records for related episodes of care * Assessment of complete submission (Comparison of two lists : hospital list of cases with specific billing codes versus NCDR submitted records) |
| **Criteria for selecting sites/records** | Remote audit :   * Sites passing their quarterly Data Quality Report for 2 quarters within audited year * Sites submitting at least the number of records/sites being reviewed   Onsite audit   * Sites identified with an outlier and not contacted with the data outlier program |
| **Scoring** | NCDR uses a grading system for identifying the amount of agreement or  matching between the data captured during the medical record review  and data submitted to the NCDR. |

**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*)  
***Signal to Noise Analysis:***

Signal to Noise analysis for the hospitals are noted under Table 5.

Table 5. Signal to Noise Analysis

|  |  |
| --- | --- |
| Level | Signal-to-Noise |
| All, >10 Procedures | .537 |
| >Q1 (>181 Procedures) | .582 |
| >Q2 (>356 Procedures) | .659 |
| >Q3 (>626 Procedures) | .748 |
| >Average (>467 Procedures) | .700 |

***Assessment of test-retest reliability among patients undergoing 2 procedures within 2012:***

The key data elements for the mortality risk model tested among patients with 2 procedures in 2012 are shown below:

**Gender** demonstrated excellent reproducibility, with only 12 of 40,197 (0.03%) patients having different genders on the 2 procedures.

**Age as assessed by Date of Birth** was identical in 99.91% of the 40,045 patients on both assessments.

**Cerebrovascular disease (CVD)** revealed that only 1160 patients had evidence of CVD on the initial visit that was not noted on the second visit. This represents 2.9% of the population being clearly misclassified on one of the assessments.

**Peripheral Vascular Disease (PVD)** revealed that only 1332 (3.3%) patients who had evidence of PVD at the time of their initial PCI no longer had this recorded at the time of their second procedure and were clearly misclassified on one of the assessments.

**Chronic Lung Disease (CLD)** was recorded in 1366 (3.4%) of the patients at the time of their initial PCI, but not at the time of the second procedure.

**Prior PCI** should have been recorded on the second procedure for each of the 40,045 patients. 987 (2.5%) were not classified as having had a prior PCI.

**Diabetes** was not recorded among 731 (1.8%) of the patients who were noted to have diabetes at the time of their original procedure.

Because dynamic elements are expected to change over time, the following variables could not have their test-retest reliability assessed by this method: Prior cardiac arrest, GFR, NYHA classification, shock within 24 hours of PCI, indication for PCI (e.g. STEMI vs. NSTEMI vs. others), urgency of the procedure, number, appearance and location of diseased vessels, lesion severity as assessed by the SCAI definitions, BMI, and TIMI flow could not be assessed using this approach.

**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?*)  
***Signal to Noise Analysis:***

The signal to noise ratio analysis measures the confidence levels in differentiating performance between hospitals. These numbers demonstrate variability that is attributable to real differences in hospital quality as opposed to measurement error.

***Assessment of test-retest reliability among patients undergoing 2 procedures within 2012:***

Finding no clear misclassification by test-retest reliability for any assessable risk factor being >3.5% provides strong support for the test-retest reliability of the mortality risk factors assessed.

Collectively, we believe that the prior audit data and repeat procedure data strongly support the reliability of the data elements used in the model.

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

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

**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)*  
**Rationale for proposing this outcome -** Peri-procedural mortality is the most dreaded complication of PCI. The currently NQF-approved risk-adjusted peri-procedural mortality model has excellent discrimination and markedly shifts the observed performance of hospitals (see Section 2b4.9 below) by accounting for patient characteristics present prior to the conduct of the procedure. Given the marked distribution of performance across hospitals, we believe that some hospitals are clearly performing PCI more safely than others and that there is great opportunity to improve the safety of PCI at some centers. Importantly, we have also created a much simpler, pre-procedural risk model that can be used clinically to assess patients’ risks for mortality. These estimates can be used by heart teams to define the best treatment strategy for each patient.

**Content validity of this outcome –**the specific definition used in defining peri-procedural mortality and the inclusion/exclusion criteria were was achieved by the specialized expertise of those individuals who developed this model as well as the structured discussions that the group conducted(Peterson et al. *J Am Coll Cardiol* 2010; 55: 1923-32). For this particular topic those individuals who were involved in identifying the key attributes and variables for this risk model were leaders and experts in the field of interventional cardiology. Serial phone calls were held to be both define the event and to examine and vet the risk model. Additional review was provided by the following specific committees and workgroups are noted below:

NCDR Strategic Quality and Oversight Committee— an ACC leadership oversight committee that serves as the primary resource for crosscutting scientific and quality of care methodological issues – ensured the data dictionaries and metrics are consistent across registries. They also reviewed and approved the methodology and results of the mortality as an outcome and model.

These members include Dr. Frederick Masoudi (chair) , Dr. David Malenka, Dr. Thomas Tsai, Dr. Matthew Reynolds,Dr. David Shahian, Dr. John Windle, Dr. Fred Resnic, Dr. John Moore, Dr. Deepak Bhatt, Dr. James Tcheng, Dr. Jeptha Curtis, Dr. Paul Chan, Dr. Matt Roe, and Dr. John Rumsfeld

NCDR Clinical SubWorkgroup is a designated set of experts that oversees this NQF application. Prior to submission, it ensures there is variation in care, disparities data, and that the measure is a true reflection of quality care at a particular site and can also be used to improve quality.

Dr. Jeptha Curtis (chair), Dr. Frederick Masoudi, Dr. John Rumsfeld, Dr. David Malenka, and Dr. Issam Moussa.

NCDR Registry Steering Committee provides strategic direction for the Registry and ensures the measures submitted to NQF met key criterion such as reliability, feasibility, and that there is compelling evidence base behind the development and implementation of this measure.

Dr. Issam D. Moussa (chair), Dr. Kirk N. Garratt, Dr. Lloyd W. Klein, Dr. Kendrick A. Shunk, Dr. Samir R. Kapadia, Dr. Robert N. Piana, Dr. Roxana Mehran, Dr. Frederic S. Resnic, Dr. Aaron D. Kugelmass,

Dr. Sunil V. Rao, Dr. W. Douglas Weaver, and Dr. John C. Messenger.

Lastly the 16 member NCDR Management Board and 31member ACCF Board of Trustees approved these measures for submission to NQF.

**2b1.3. What were the statistical results from validity testing**? (*e.g., correlation; t-test*)  
No validity testing was necessary, other than establishing the content validity of the model, as mortality is of unquestioned importance and readily assessed.

**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?*)  
We believe that the outcome is of clear importance and the construct of the risk-adjustment model has been thoroughly vetted and published in the peer-reviewed literature. Prior endorsement by NQF further supports the logic and care we used in developing this performance measure.

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

**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*)  
 The only exclusion is for patients transferred to another acute care facility, in whom their vital status cannot be readily determined.

**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*)  
There were 8,619 (<1%) patients transferred to another acute facility.

**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*)  
We do not believe that the exclusions have any impact on the validity, accuracy or interpretability of the risk-adjusted in-hospital mortality outcome measure.

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

**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** 40 **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**.   
This is not relevant as we are proposing a risk-adjusted peri-procedural mortality outcome measure to help assess the quality and safety of PCI.

**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?  
As described in Section 2b.1.2, there was an extensive process to develop the face and contact validity of the measure. After settling on the outcome definition and candidate variables through serial conference calls with the expert panel, categorical variables were summarized as frequencies and percentages and compared with Pearson chi-squared tests. Continuous variables were summarized as medians (interquartile range) and compared using Wilcoxon rank-sum tests. Ordinal variables were tested using a chi-square test based on the rank of the group mean score.

The original model was developed by using a sample of 181,775 NCDR patients undergoing PCI from 1/04-3/06 and then validated in 2 separate samples; an additional 121,183 patients treated in the same time period and a prospective cohort of 285,440 patients treated between 3/06-3/07. Baseline patient characteristics and variables from diagnostic catheterization were considered candidate variables. Candidate variables had less than 0.5% missing data except for pre-procedure ejection fraction (29.7%). Missing values for ejection fraction were imputed by stratifying the population based on a history of congestive heart failure, prior MI, pre-procedural cardiogenic shock and the presence of STEMI to determine a median value for each patient with missing data. After the committee reviewed all variables with a statistically significant univariate association with mortality, the most clinically and statistically meaningful values were selected for potential inclusion in a logistic regression model. Backward selection with a ‘stay’ criterion of p<0.05 to develop a model predicting post-PCI mortality was then created. Variables that showed non-linear associations with the outcome were transformed using splines. All 2-way interactions were examined and significant ones were retained.

We also developed a simplified pre-procedural risk model by relying only upon pre-catheterization data that had the strongest association with mortality. This simplified model had similar discrimination and calibration and enables clinicians to estimate patients’ peri-procedural mortality and share this information with patients and use the estimates to define the safest and best care for each individual patient.

The C-statistic was used to describe the discrimination of the model and replicated in clinically important subgroups of interest, including patients with and without STEMI, males and females, those aged > and ≤70 years, and patients with and without diabetes. Calibration plots were used to access goodness of fit. A p-value <0.05 was considered statistically significant. All statistical tests were two-sided. All statistical analyses were performed using SAS software (version 9.2, SAS Institute, Cary, NC).

**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)**

Social risk factors were not used in the risk modeling

**2b3.4a. What were the statistical results of the analyses used to select risk factors?**As described above, bivariate analyses were done to identify candidate variables that differed significantly between those with and without a clinically important mortality event. Multivariable, logistic regression analyses were then performed to retain those clinically meaningful variables with a statistically significant association with mortality (p<0.05 for each). Table 3 in Section 1.6 demonstrates the difference between those who did and did not die after their procedure, based upon 2012 data.

**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*)  
The process for developing the model is described in section 2b3.3 above. Discrimination was assessed with the c-statistic and calibration was assessed with both the Hosmer-Lemeshow test and the slope of the predicted vs. observed risk. Given that the prior, approved submission included the results for the separate derivation and validation cohorts reported in Peterson et al (*J Am Coll Cardiol* 2010; 55: 1923-32), we report only the 2012 data here.

*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*)**:**The c-statistic is 0.93, which means that the probability that predicting the outcome is substantially better than chance. This method is used to compare the goodness of fit of logistic regression models. The range is between 0.5 to 1.0. A value of 0.5 indicates that the model is no better than chance at making a prediction of membership in a group and a value of 1.0 indicates that the model perfectly identifies those within a group and those not. Models are typically considered reasonable when the C-statistic is higher than 0.7. (Hosmer & Lemeshow, 2000).

The c-statistics for the original derivation and validation cohorts, as well as clinically important subgroups are provided under Table 6.

Table 6. C-Statistic Results

|  | **Sample, n** | **Full Model C-Stats** |
| --- | --- | --- |
| Development | 181,775 | 0.926 |
| 1st validation | 121,183 | 0.925 |
| 2nd validation | 285,440 | 0.924 |
| Subgroups (in 2nd validation) |  |  |
| STEMI | 39,889 | 0.902 |
| No STEMI | 245,551 | 0.892 |
| Women | 95,106 | 0.911 |
| Men | 190,334 | 0.930 |
| Age >70 yrs | 92,381 | 0.901 |
| Age ≤70 yrs | 193,059 | 0.927 |
| Diabetes | 92,974 | 0.924 |
| No diabetes | 192,466 | 0.923 |

Cath = catheterization; NCDR = National Cardiovascular Data Registry; STEMI = ST-segment elevation myocardial infarction.

**2b3.7. Statistical Risk Model Calibration Statistics** (*e.g., Hosmer-Lemeshow statistic*):   
The intercept for the model was -0.00063, which was not statistically significantly different than 0 (p=0.97). The slope of the calibration line was 0.9906, which also was not significantly different than 1.0 (p=0.097). A graphical representation of observed and predicted mortality rates across deciles of risk is shown under Figure 1.

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


Figure 1. Calibration Curve Plot

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

The risk stratification was able to adequately segregate deciles of risk from <1% to >12% at the patient level. At the hospital level, we observed a broad range of unadjusted risk, which was substantially tightened after adjusting for patient characteristics. The unadjusted distribution of mortality is shown under Figure 2.

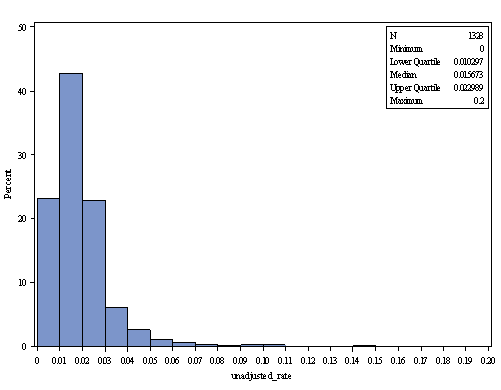


Figure 2. Unadjusted Distribution of Mortality

The mortality rates adjusted for patient characteristics is shown under Figure 3.



Figure 3. Adjusted Distribution of Mortality

After adjusting for patient characteristics, we observed a significantly tighter and more normal distribution of mortality events.

The distribution of sites’ observed/expected ratios are shown under Figure 4.



Figure 4. Site’s Observed and Expected Ratio

**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*)  
We believe that the our mortality model performs exceedingly well in adjusting for patient characteristics present prior to the conduct of PCI and is able to discriminate well across a wide variety of important clinical subsets of patients. Moreover, there is substantial hospital variation before and after risk-adjusting patient characteristics. The distribution of hospitals’ O/E ratios show that there are some sites with excellent performance and others with mortality rates that are more than 2-fold greater than expected. These would be sites where substantial opportunities to improve patient safety likely exist.

**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*)

Since these data provide further strong evidence of the validity and value of the previously-endorsed measure using 2012 data, we did not do any additional testing.

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

**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)*   
 As noted in the figures above, we found significant variability in mortality across hospitals, even after adjusting for pre-procedural patient characteristics. Those in the upper quartile of performance had an observed/expected ratio that was 31% greater than predicted, with some sites having a greater than 5-fold excess mortality over that predicted.

**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*)  
A meaningful difference is one that indicates the potential for improvement in comparison to others. There are no absolute levels of mortality that are significant as compared with others. The average, adjusted mortality rate was 1.6% and the upper quartile ranges from 2.1 to 14%. Given an average PCI volume of 410 cases/hospital, this suggests between 2 and 50 extra deaths might be avoided per year among hospitals in the upper quartile as compared with the average hospital. Clinically, this is a large number of events, and the few significant outliers would have a very strong incentive to improve the safety of their PCI procedures.

**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?*)  
We believe that the use of this model to identify outliers and the ability to pre-procedurally risk stratify patients and tailor therapy to risk holds great promise for improving the quality and safety of PCI.

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

**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.***

**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*)

**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*)

**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*)