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

**Measure Number** (*if previously endorsed*)**:** 2459

**Measure Title**: In-hospital Risk Adjusted Rate of Bleeding Events for patients undergoing PC

**Date of Submission**: 08/01/2018

**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 decision-making) 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 used the National Cardiovascular Data Registry for CathPCI Registry. This is a national quality improvement registry with more than 1200 participating US hospitals. Participation is largely voluntary though 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

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

1. Creation of the Bleeding model was performed on all national NCDR data from 02/2008–04/2011 and has been used to provide a description and initial performance characteristics of the model.

2. A validation cohort from the NCDR CathPCI was identified (all cases performed between 01/2016-12/2016). These data were also used to assess test-retest reliability of the risk model covariates and validate the association between the predictor variables and bleeding, including model discrimination and calibration.

**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 Bleeding & Validation model:*

The model was originally developed using data from 1,142 hospitals. See additional information under section 1.6.

*Test-Retest*

The 2016 validation sample includes cases from 1,619 hospitals.

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

*Bleeding & Validation Model:*

For the initial derivation and validation of the bleeding risk model, 1,043,759 patients undergoing PCI between 2/2008-4/2011 at 1,142 hospitals were included; 80% were randomly assigned to the derivation cohort with the remaining 20% serving as the validation cohort. Of these, 60,194 PCI procedures had post-procedure bleeding, yielding a post-PCI bleeding event rate of 5.8%. A summary of these patients’ clinical characteristics and the hospital characteristics are provided under Table 1:

Table 1. Derivation and Validation Characteristics

|  | **Overall (n=1,043,759)** | **Development (n=834,696)** | **Validation**  **(n=209,063)** |
| --- | --- | --- | --- |
| Demographics |  |  |  |
| Median age, y (25th, 75th percentiles) | 65.0  (56.0, 74.0) | 64.0  (56.0, 74.0) | 65.0  (56.0, 74.0) |
| Female sex | 32.7 | 32.6 | 32.8 |
| Median BMI, kg/m2  (25th, 75th percentiles) | 29.1  (25.7, 33.3) | 29.1  (25.7, 33.3) | 29.1  (25.7, 33.3) |
| Medical conditions |  |  |  |
| Diabetes mellitus | 35.9 | 35.9 | 35.9 |
| Hypertension | 81.8 | 81.8 | 81.9 |
| Peripheral vascular disease | 12.4 | 12.4 | 12.4 |
| Chronic kidney disease | 3.6 | 3.6 | 3.6 |
| Prior PCI | 40.3 | 40.3 | 40.3 |
| Prior CABG | 18.8 | 18.9 | 18.7 |
| Median pre-procedure Hgb, g/dl  (25th, 75th percentiles) | 13.7  (12.4, 14.9) | 13.7  (12.4, 14.9) | 13.7  (12.4, 14.9) |
| Procedural characteristics |  |  |  |
| Procedure status |  |  |  |
| Elective | 45.2 | 45.2 | 45.1 |
| Urgent | 37.5 | 37.5 | 37.7 |
| Emergent | 17.0 | 17.0 | 16.9 |
| Salvage | 0.3 | 0.3 | 0.3 |
| STEMI | 16.0 | 16.0 | 15.9 |
| Lytics prior to PCI for STEMI | 8.1 | 8.0 | 8.2 |
| Shock | 2.5 | 2.5 | 2.4 |
| Cardiac arrest within 24 hrs of PCI | 1.7 | 1.7 | 1.7 |
| Hospital characteristics |  |  |  |
| Number of beds, median  (25th, 75th percentiles) | 410.0  (283.0-571.0) | 410.0  (283.0-571.0) | 409.0  (282.0-569.0) |
| University hospital (%) | 11.3 | 11.3 | 11.3 |
| Number of annual PCI cases, median  (25th, 75th percentiles) | 726.0  (445.1-1177.9) | 726.6  (445.1-1183.1) | 726.6  (448.0-1177.9) |
| All p-values >0.05  BMI = body mass index; CABG = coronary artery bypass grafting; Hgb = hemoglobin; PCI = percutaneous coronary intervention; STEMI = ST-segment elevation myocardial infarction | | | |

*Test-Retest*

To test the predictive validity, calibration and test-retest reliability, we used data from 717,510 patients undergoing PCI between 1/2016-12/2016, of whom 23,874 (3.3%) had a bleeding event. A summary of these patients’ clinical characteristics are provided in Table 2 below:

Table 2. Predicted Probability of Bleeding (2016)

| Bleeding | | | |
| --- | --- | --- | --- |
|  | Total | Observed Bleed | |
| n = 717510 | Bleed n = 23874 | No Bleed n = 693636 |
| Predicted Bleeding Risk using the proposed risk-adjustment model (multiply by 100 to get % bleeding risk) | 0.03707 ± 0.05064 | 0.10379 ± 0.11061 | 0.03477 ± 0.04553 |
| ***Bleeding Variables*** |  |  |  |
| STEMI | 121466 (16.9%) | 9699 (40.6%) | 111767 (16.1%) |
| Age | 65.6 ± 11.9 | 68.4 ± 12.5 | 65.5 ± 11.9 |
| BMI | 30.2 ± 6.5 | 28.9 ± 6.8 | 30.2 ± 6.5 |
| CVD | 96012 (13.4%) | 4300 (18.0%) | 91712 (13.2%) |
| PVD | 88106 (12.3%) | 4008 (16.8%) | 84098 (12.1%) |
| CLD | 111555 (15.5%) | 4956 (20.8%) | 106599 (15.4%) |
| Prior PCI | 295177 (41.1%) | 7482 (31.3%) | 287695 (41.5%) |
| Insulin Diabetes | 116320 (16.2%) | 4540 (19.0%) | 111780 (16.1%) |
| Dialysis | 21673 (3.0%) | 1591 (6.7%) | 20082 (2.9%) |
| GFR 45-60 (Mild) | 102839 (14.3%) | 4592 (19.2%) | 98247 (14.2%) |
| GFR 30-45 (Moderate) | 58134 (8.1%) | 4779 (20.0%) | 53355 (7.7%) |
| Lytics Prior to PCI for STEMI | 5041 (0.7%) | 366 (1.5%) | 4675 (0.7%) |
| Cardiogenic Shock at Start of PCI           Missing | 16695 (2.3%) 332 | 4227 (17.7%) 8 | 12468 (1.8%) 324 |
| Cardiogenic Shock w/in 24 Hours           Missing | 14743 (2.1%) 85 | 4080 (17.1%) 2 | 10663 (1.5%) 83 |
| PCI Status                      Elective                      Urgent                      Emergent                      Salvage           Missing | 251808 (35.1%) 331775 (46.3%) 131388 (18.3%) 2359 (0.3%) 180 | 3572 (15.0%) 9161 (38.4%) 10331 (43.3%) 806 (3.4%) 4 | 248236 (35.8%) 322614 (46.5%) 121057 (17.5%) 1553 (0.2%) 176 |
| In-stent Thrombosis | 2420 (0.3%) | 245 (1.0%) | 2175 (0.3%) |
| Lesion SCAI Class II/III | 352960 (49.2%) | 11240 (47.1%) | 341720 (49.3%) |
| Lesion SCAI Class IV | 107008 (14.9%) | 7304 (30.6%) | 99704 (14.4%) |
| Prox LAD | 129513 (18.1%) | 5251 (22.0%) | 124262 (17.9%) |
| Left Main | 20218 (2.8%) | 1780 (7.5%) | 18438 (2.7%) |
| HF NYHA Class IV | 24810 (3.5%) | 3171 (13.3%) | 21639 (3.1%) |
| HF NYHA Class I/II/III | 75463 (10.5%) | 3456 (14.5%) | 72007 (10.4%) |
| Cardiac arrest w/in 24 hrs | 15060 (2.1%) | 3190 (13.4%) | 11870 (1.7%) |
| Lesion: Preprocedure TIMI Flow = NO | 141133 (19.7%) | 9191 (38.5%) | 131942 (19.0%) |
| Multivessel Disease | 297871 (41.5%) | 12302 (51.5%) | 285569 (41.2%) |
| PreHGB | 13.5 ± 2.0 | 12.7 ± 2.6 | 13.5 ± 2.0 |
| Female | 223562 (31.2%) | 10961 (45.9%) | 212601 (30.7%) |
| ***History*** |  |  |  |
| IABP           Missing | 12563 (1.8%) 194 | 3207 (13.4%) 4 | 9356 (1.3%) 190 |
| Current/Recent Smoker (w/in 1 year)           Missing | 182938 (25.5%) 309 | 6401 (26.8%) 16 | 176537 (25.5%) 293 |
| Hypertension           Missing | 598736 (83.5%) 136 | 19532 (81.8%) 8 | 579204 (83.5%) 128 |
| Dyslipidemia           Missing | 558972 (78.0%) 449 | 16761 (70.3%) 28 | 542211 (78.2%) 421 |
| Family History of Premature CAD           Missing | 133606 (18.6%) 330 | 3101 (13.0%) 21 | 130505 (18.8%) 309 |
| Prior MI           Missing | 218401 (30.4%) 171 | 6888 (28.9%) 9 | 211513 (30.5%) 162 |
| Prior Heart Failure           Missing | 111292 (15.5%) 206 | 5641 (23.6%) 12 | 105651 (15.2%) 194 |
| Prior Valve Surgery/Procedure           Missing | 13268 (1.8%) 279 | 620 (2.6%) 14 | 12648 (1.8%) 265 |
| Prior CABG           Missing | 124219 (17.3%) 79 | 3451 (14.5%) 3 | 120768 (17.4%) 76 |
| ***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 | 110008 (15.3%) 7334 (1.0%) 1556 (0.2%) 1556 (0.2%) 3060 (0.4%) 426385 (59.4%) 34899 (4.9%) 132555 (18.5%) 157 | 8588 (36.0%) 820 (3.4%) 103 (0.4%) 43 (0.2%) 261 (1.1%) 11019 (46.2%) 744 (3.1%) 2295 (9.6%) 1 | 101420 (14.6%) 6514 (0.9%) 1453 (0.2%) 1513 (0.2%) 2799 (0.4%) 415366 (59.9%) 34155 (4.9%) 130260 (18.8%) 156 |
| CAD Presentation           No symptom, no angina           Symptom unlikely to be ischemic           Stable angina           Unstable angina           Non-STEMI           ST-Elevation MI (STEMI) or equivalent           Missing | 26571 (3.7%) 13214 (1.8%) 91567 (12.8%) 283204 (39.5%) 181336 (25.3%) 121466 (16.9%) 152 | 755 (3.2%) 424 (1.8%) 1291 (5.4%) 4932 (20.7%) 6767 (28.4%) 9699 (40.6%) 6 | 25816 (3.7%) 12790 (1.8%) 90276 (13.0%) 278272 (40.1%) 174569 (25.2%) 111767 (16.1%) 146 |
| Anginal Classification w/in 2 Weeks           No symptoms           CCS I           CCS II           CCS III           CCS IV           Missing | 51063 (7.1%) 13240 (1.8%) 74955 (10.5%) 275908 (38.5%) 300897 (42.0%) 1447 | 2651 (11.1%) 274 (1.2%) 1214 (5.1%) 5622 (23.6%) 14049 (59.0%) 64 | 48412 (7.0%) 12966 (1.9%) 73741 (10.7%) 270286 (39.0%) 286848 (41.4%) 1383 |
| Anti-Anginal Medication w/in 2 Weeks           Missing | 532589 (74.2%) 213 | 15753 (66.0%) 11 | 516836 (74.5%) 202 |
| Heart Failure w/in 2 Weeks           Missing | 100273 (14.0%) 290 | 6627 (27.8%) 7 | 93646 (13.5%) 283 |
| Cardiomyopathy or Left Ventricular Systolic Dysfunction           Missing | 91302 (12.7%) 111 | 4704 (19.7%) 2 | 86598 (12.5%) 109 |
| Pre-operative Evaluation Before Non-Cardiac Surgery           Missing | 13398 (1.9%) 107 | 386 (1.6%) 3 | 13012 (1.9%) 104 |
| Pre-PCI Left Ventricular Ejection Fraction           Missing | 52.0 ± 12.8 216496 | 45.5 ± 15.7 9862 | 52.2 ± 12.6 206634 |
| ***Procedure Information*** |  |  |  |
| Contrast Volume           Missing | 179.9 ± 81.4 3382 | 198.7 ± 99.2 149 | 179.2 ± 80.6 3233 |
| Fluoroscopy Time           Missing | 15.8 ± 12.4 15939 | 21.4 ± 18.8 644 | 15.6 ± 12.0 15295 |
| ***Outcomes*** |  |  |  |
| Discharge Status           Alive           Deceased | 707874 (98.7%) 9636 (1.3%) | 20522 (86.0%) 3352 (14.0%) | 687352 (99.1%) 6284 (0.9%) |
| Primary Cause of Death           Cardiac           Neurologic           Renal           Vascular           Infection           Valvular           Pulmonary           Unknown           Other           Missing | 6948 (72.2%) 642 (6.7%) 79 (0.8%) 93 (1.0%) 274 (2.8%) 206 (2.1%) 535 (5.6%) 366 (3.8%) 477 (5.0%) 16 | 2323 (69.5%) 241 (7.2%) 23 (0.7%) 61 (1.8%) 122 (3.6%) 73 (2.2%) 149 (4.5%) 122 (3.6%) 230 (6.9%) 8 | 4625 (73.7%) 401 (6.4%) 56 (0.9%) 32 (0.5%) 152 (2.4%) 133 (2.1%) 386 (6.2%) 244 (3.9%) 247 (3.9%) 8 |
| Myocardial Infarction (Biomarker Positive)           Missing | 10766 (1.5%) 26 | 1016 (4.3%) 6 | 9750 (1.4%) 20 |
| Cardiogenic Shock           Missing | 9697 (1.4%) 7 | 3414 (14.3%) 1 | 6283 (0.9%) 6 |
| Heart Failure           Missing | 10082 (1.4%) 8 | 2304 (9.7%) 2 | 7778 (1.1%) 6 |
| CVA/Stroke           Missing | 2164 (0.3%) 11 | 759 (3.2%) 4 | 1405 (0.2%) 7 |
| Other Vascular Complications Requiring Treatment           Missing | 2285 (0.3%) 14 | 1143 (4.8%) 1 | 1142 (0.2%) 13 |
| RBC/Whole Blood Transfusion | 12641 (1.8%) | 11297 (47.3%) | 1344 (0.2%) |
| Continuous variables compared using Student's T-test. Categorical variables compared using chi-square or Fisher's exact test.  **All p-values were <0.001** | | | |

**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**.

As noted above, the ACCF used different data sources for different analyses. The original model was developed and validated using data from 1,043,759 patients undergoing PCI between 2/2008-4/2011 at 1,142 hospitals.

A reassessment of model performance was performed using all PCI patients enrolled (n=715,510) in the NCDR CathPCI registry in calendar year 2016. Separately, we identified 42,637patients who underwent 2 PCIs within the 2016 calendar year in whom we were able to assess the test-retest reliability of the data elements used to predict patients’ bleeding risks.

**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 clincial registry used for quality assessment and improvement, detailed socioeconomic variables are not available. Second, while proxy variables could be considered, these were not felt to be relevant to an inpatient bleeding 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. clinical indication for PCI, Hb, etc.) is a much more accurate means of stratifying risk. Accordingly, we feel that in this model of in-hospital risk-adjusted bleeding rate, given the rich clinical data available through the NCDR CathPCI registry, that social risk factors, which are not readily available, would not likely improve 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*)

Performance Measure Score (Signal-to-Noise):

ACCF performed the signal-to-noise analysis on the same cohort of individuals as noted under Section 1.3. (testing method 2). For the signal-to-noise analysis, we followed the methodology as outlined in a Rand Corporation technical report by John L Adams. The document is available at the following URL (<https://www.rand.org/content/dam/rand/pubs/technical_reports/2009/RAND_TR653.pdf>).  This approach uses a beta-binomial model that assumes the physician’s score is a binomial random variable conditional on the physician’s true value that comes from a beta distribution. The beta distribution is a very flexible distribution on the interval from 0 to 1 and can have any mean within the interval and can be skewed left or right or even U-shaped. It is the most common distribution for probabilities on the 0-1 interval. Signal to Noise analysis for the hospitals participating in 2016 are provided in Table 3. The author used a beta-binomial model, specifically the Betabin SAS macro to output the required parameters in the reliability formula provided.

Data Element (Test-Retest Reliability):

ACCF evaluated the test-retest reliability by reviewing CathPCI patients who were readmitted or had a repeat procedure in 2016. 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.

***Signal to Noise Analysis:***

Signal to Noise analysis for the hospitals participating in 2016 are provided in Table 4.

Table 3. Signal to Noise Analysis

|  |  |
| --- | --- |
| Level | Signal-to-Noise |
| All Procedures | .743 |
| >Q1 (>185 Procedures) | .706 |
| >Q2 (>360 Procedures) | .760 |
| >Q3 (>628 Procedures) | .819 |
| >Average (>470 Procedures) | .791 |

***Assessment of test-retest reliability among patients undergoing 2 procedures within 2016:***

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

**Gender** demonstrated excellent reproducibility, with only 18 of 42,637 (0.06%) patients having different genders on the 2 procedures.

**Age as assessed by Date of Birth** was identical in 99.90% of the 42,637 patients on both assessments.

**Cerebrovascular disease (CVD)** revealed that only 1213 patients had evidence of CVD on the initial visit that was not noted on the second visit. This represents a 2.84% misclassification rate for one of the assessments.

**Peripheral Vascular Disease (PVD)** revealed that only 1282 (3.0%) 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. This represents a 3.0% misclassification rate for one of the assessments.

**Chronic Lung Disease (CLD)** was recorded in 1,294 (3.0%) 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 42,637 patients. 1259 (2.95%) were not classified as having had a prior PCI.

**Diabetes** was not recorded among 745(1.75%) 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 test-retest reliability of the following could not be assessed by this method: Prior cardiac arrest, GFR, NYHA classification, shock within 24 hours of PCI, indication for PCI, urgency of the procedure, use of fibrinolysis prior to PCI, pre- and post-procedure hemoglobin, number and location of diseased vessels, lesion severity as assessed by the SCAI definitions, pre-procedural TIMI flow and acute stent thrombosis.

**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 2016:***

Finding no clear misclassification by test-retest reliability for any assessable risk factor being >3.0% provides strong support for the test-retest reliability of the bleeding risk factors assessed.

Collectively, we believe that the test-retest reliability data and signal to noise analysis strongly support the reliability of the data elements and measurement scores used in the model.

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

**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)*

We performed 2 different strategies for assessing the validity of this measure. First, we underwent a rigorous process for establishing the face validity of the measure. Because it is a clinically meaningful outcome, we sought to make sure that a broad range of experts and clinicians concurred that this was a clinically important outcome measure. Second, we hypothesized that it would be associated with other clinically important outcomes and sought to establish the predictive validity of the measure. These are described in more detail below:

Systematic Assessment of Face Validity of the Performance Measure:

Bleeding remains one of the most common non-cardiac complications of PCI. It is a serious adverse consequence and, most importantly, is modifiable. The 2011 ACC/AHA guidelines provide for a Level IC recommendation for the assessment of bleeding prior to PCI. This is grounded in the realization that there are several strategies, such as radial approaches and the use of bivalirudin, that can be applied to mitigate the risk of bleeding, particularly in high-risk patients. The first bleeding risk model was published in 2009 (*Circ Cardiovasc Intervent*. 2009;2:222-229) and was the update was published in 2013 (*JACC Cardiovasc Intervention, 2013;6:897-904).*

Content validity of this outcome – and the specific definition used in defining a bleeding event – was achieved by the specialized expertise of those individuals who developed this model as well as the structured discussions that the group conducted. 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. Multiple conference calls were held to both define a bleeding event and to examine and vet the risk model. These individuals within specific committees and workgroups are noted below:

NCDR Science and Quality 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 bleeding outcome and model.

These members include

John C. Messenger, MD, FACC (Chair); David M. Shahian, MD, FACC; Thomas T Tsai, MD, MSC; Charles A. Henrikson, MD, MPH; Jeff Jacobs, MD, FACC; John R. Windle, MD, FACC; Amit Amin, MD; John W. M. Moore, MD, FACC; Deepak L. Bhatt, MD, MPH, FACC; Jeffrey Westcott, MD, FACC; Gregory M. Marcus, MD FACC; David J. Slotwiner, MD, FACC; Jeptha P. Curtis, MD, FACC; John Spertus, MD, FACC; Matthew T. Roe, MD, FACC; and Frederick A. Masoudi, MD, MSPH, FACC

NCDR Clinical SubWorkgroup was a designated workgroup that oversaw the initial NQF application. Prior to submission, the group ensured there was 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 provided 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.

The NCDR Metrics and Reporting Methodology (MRM) Subcommittee of the Science and Quality Oversight Committee, reviews for re-endorsement and a data analytic center is involved in evaluating data, providing corresponding analysis/interpretation of data. The review includes guidance and oversight from both NCDR’s Chief Science Officer (Frederick Masoudi) and chair of MRM (Jeptha Curtis).

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

In addition, the NCDR provides an open comment period (typically between 15 and 30 days) for: 1) all registry data set version changes, 2) new registry version measures and 3) significant changes/additions to registry version metrics/measures, including risk models and appropriate use criteria. The open comment period engages key registry shareholders (i.e., physicians and clinical care team members and hospital or practice representatives) as well as other external stakeholders (i.e., hospitals, physicians, payers, regulators, consumers, purchasers, etc.) Comments submitted are considered for modification of the version change. NCDR staff and members involved in developing the measures and reports receive all the comments submitted including the name of the individual and organization submitting comment. The NCDR determines which comments to incorporate into modifications and the internal timeline for any modifications. No formal response is provided back to individuals submitting comments through this process. The NCDR may choose to provide a report of comments received and decisions made regarding the various feedback to a broader audience.

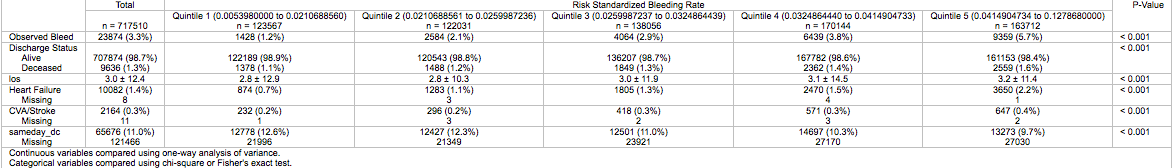
Beyond the inherent content validity of this process, we have data showing that the bleeding risk score is highly actionable – a critical feature for moving beyond quality assessment to quality improvement. For example, a comparative effectiveness analysis of bivaluridin use by bleeding risk suggested that bivalirudin was preferentially used in low-risk patients (NNT=224) and least often used in patients at high risk for bleeding (NNT=43; *JAMA* 2010;303(21):2156-2164). At Saint Luke’s Mid America Heart Institute, the original bleeding model was executed prior to non-emergent PCI in all patients undergoing the procedure. Not only was the ‘risk-treatment’ paradox reversed, but the bleeding rate at that institution decreased by 40% (*J Am Coll Cardiol* 2013;61: 1847–52). More recently, a 9-center study of providing pre-procedural bleeding risks demonstrated a fully-adjusted 44% lower odds of bleeding when the models were used (*BMJ*, 2015;350:h1302). The ultimate validity of the model is that the use of the model to target therapy improves outcomes strongly supports the appropriateness and capacity of this model to measure and improve quality.

Predictive Validity:

To further underscore the importance of the bleeding measure, we examined the association of bleeding rates, by quintiles, with other clinically important outcomes, including mortality, complications of heart failure and stroke, length of stay and rates of same-day discharge. We hypothesized that patients experiencing a bleeding complication would also be at higher risk for longer post-procedure lengths of stay (because additional observation and treatment, such as transfusions and surgical repairs, would be needed to address the bleeding complication. Other complications that we hypothesized to be associated with bleeding events would be a greater risk for other complications, including stroke, heart failure and mortality. We postulated that the anemia and hypotension associated with severe bleeds could put patients at increased risk for stroke and death, while the fluid resuscitation and transfusions used to treat a bleeding event might be associated with heart failure exacerbations during the hospitalization. Because it is a hospital-based measure, we examined the risks of these other adverse outcomes across quintiles of bleeding rates throughout the NCDR registry. For the importance/predictive validity of this measure, we found important associations with all outcomes:

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

Predictive Validity:



The above table shows quintiles of performance based on hospitals’ risk standardized bleeding rate from best performing (i.e. Quintile #1 – bleeding rate range of 0.5-2.1%) to worst performing hospitals (i.e. Quintile #2 – bleeding rate range of 4.1 to 12.8%) compared to other adverse event measures. For example, if mortality observed in the best performing hospitals had been observed in the worst performing hospitals, then 1826 deaths would have been averted. Similar patterns were observed for the complications of heart failure (3-fold increase in risk of heart failure from worst to best performing hospitals) and stroke (a doubling of risk from the hospitals with the lowest vs. the highest rates of bleeding). There was also evidence of more efficient care, with an observed LOS of 2.8 vs. 3.2 days and a same-day discharge rate of 12.6 vs. 9.7% between the best and worst performing hospitals. All of these associations support our hypotheses that bleeding, and its treatment, would be associated with other clinically important outcomes and support the predictive validity of the proposed bleeding measure.

**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?*)

Face-Validity:

As described above (2b1.2), we undertook an extensive effort to establish the definition and utility of risk-adjusted bleeding as a quality metric. These included an expert team developing the model, a group of experts, the Strategic Oversight Committee, overseeing the work and reporting of the measure – including ascertaining its alignment with both ACC/AHA PCI Guidelines and the Society of Coronary Angiography and Intervention’s (SCAI’s) 2016 Expert Consensus Statement – and an NCDR Oversight Group for NQF measures. It further underwent public comment and approval by the NCDR Management Board of the ACC’s Board of Trustees. Beyond these traditional ascertainments of its face validity, we further leveraged evidence that the prospective use of the model was associated with a substantial reduction in bleeding after PCI, clearly demonstrating the model to serve as a means for improving the safety of PCI.

Predictive Validity:

The predictive validity of risk-adjusted bleeding being associated with mortality, post-PCI complications (stroke and heart failure exacerbations) and length of stay strongly underscores the importance of this adverse event and supported the hypothesized associations in conducting these analyses. Moreover, the actionability of the bleeding model suggests that bleeding rates can be improved by prospectively using this risk-adjustment model (*BMJ*, 2015;350:h1302). Given the broad range of bleeding outcomes in the US (range of unadjusted peri-procedural bleeding in 2016 across hospitals = 0-13%, with the adjusted rate 10th to 90th percentiles of hospitals bleeding rates = 1.7-5.0%) the use of this model to assess quality and inspire improvement is a critical step towards greater patient safety and outcomes. The fact that we have demonstrated that the model, when employed at the hospital level, can reduce the variation to those factors most under the locus of control of the operators/hospitals and that by providing pre-procedural risk estimates to providers we can improve the rational use of bleeding avoidance therapies and lower bleeding strongly support the validity of this performance measure not only in risk-adjusting bleeding, but also in improving care.

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

**2b2. EXCLUSIONS ANALYSIS**

**NA**  **no exclusions — *skip to section*** [***2b3***](#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 exclusions from the bleeding model are patients undergoing CABG surgery, those who present to the hospital severely anemic (pre-procedure hemoglobin <8 g/dL) and do not have an obvious clinical bleed after their procedure, and patients who died during hospitalization (i.e. same day as their PCI procedure). These exclusions are relatively rare and firmly supported by the clinical rationale that a) bleeding and blood transfusions are common after cardiopulmonary bypass surgery and not necessarily related to the safety and quality of the PCI procedure; and b) that patients presenting to the hospital with severe anemia and receiving a blood transfusion may have been likely to be treated with a blood transfusion had they not undergone PCI. Lastly, patients who died at hospitalization would be captured under NQF measure 0133 which ACC believes complements this measure to ensure good care for patients who undergoing PCI.

**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*)

In 2016:

* CABG during the index hospitalization was performed in 8,137 (1/1%) patients. Beyond the small size of this excluded group, bleeding after a major operation is much more likely related to the operation than the preceding PCI.
* In 2016, 3,534 (0.47%) patients were excluded because they died the same day as their PCI procedure. They were excluded because they were not alive long enough to assess whether or not a bleeding event had occurred.
* There were 1,816 (0.25%) procedures in 2016 for which patients both had pre-procedural anemia (hgb<=8) and a transfusion. Of these, 1,344 (0.18%) had no bleeding evidence, and 472 had a bleed that was counted in the numerator of the bleeding measure. We do not believe that such a small rate for this exclusion would meaningfully impact the measure.
* In 2016, there were 22,406 procedures were excluded for being a second, non-index. Because it was not possible to separate which procedure, when multiple were performed in the same admission, was responsible for the bleeding event, the measure could be interpreted as bleeding events per admission in which PCI is performed, which is more clinically interpretable,

**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 bleeding 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*** [***2b4***](#section2b5)***.***

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

**No risk adjustment or stratification**

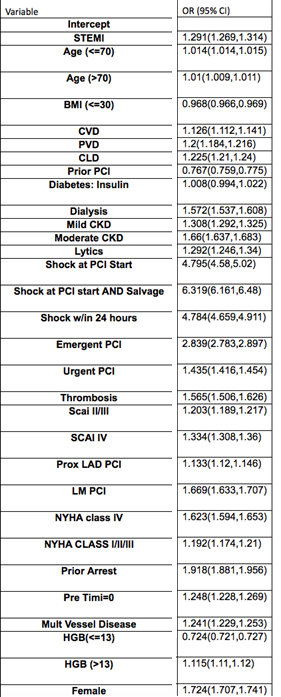
**Statistical risk model with** 32 **risk factors**

**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.**

A hierarchical logistic regression model was created. The data definitions are available on the NCDR website (<https://cvquality.acc.org/NCDR-Home/registries/hospital-registries/cathpci-registry>). The beta coefficients and covariance matrix are available from NCDR upon request.



**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?

As described in Section 1.8, we did not believe that social factors needed to be included in risk-adjusting outcomes for peri-procedural bleeding after PCI. This was predicated on the feasibility (and current unavailability) of patient-level social factors. The belief that the consequence of adverse social factors (e.g. leading to greater rates of obesity, hypertension, smoking or other comorbidities) would be directly captured by our rich clinical data, and that the short duration of follow-up (72 hours, during which the patient was hospitalized), would negate potential barriers to healthcare access and treatment that might be more relevant with longer-term outcomes. Accordingly, we feel that in this model of in-hospital risk-adjusted bleeding rate, given the rich clinical data available through the NCDR CathPCI registry, that social risk factors, which are not readily available, would not likely improve this particular risk model.

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 study population was then randomly split into a development sample consisting of 80% of PCI procedures and a validation sample consisting of the remaining 20% of admissions. Baseline patient characteristics and variables from diagnostic catheterization were considered candidate variables. Candidate variables had less than 0.5% missing data except for estimated glomerular filtration rate (7.8%), pre-procedure hemoglobin (9.5%), and ejection fraction (29.4%). Missing values were imputed to the lower risk group for discrete variables and replaced with gender-specific medians for body mass index (BMI), gender and renal failure/dialysis-specific medians for estimated glomerular filtration rate, median value for hemoglobin, and congestive heart failure (CHF)/cardiogenic shock/prior myocardial infarction (MI)-specific medians for ejection fraction. We used logistic regression with backward selection with a ‘stay’ criterion of p<0.05 to develop a model predicting post-PCI bleeding. Variables that showed non-linear associations with the outcome were transformed using splines.

We developed a full post-PCI bleeding model using all potential predictive variables. A logistic regression model with backward selection and a retention criterion of p<0.05 was performed to develop the full risk model used for hospital comparisons. Of note, a more parsimonious model for clinical use was also developed by only using those variables with the strongest association (F-statistic >500) To further simplify prospective application of the simplified model the regression coefficients from the pre-procedure model were assigned an integer that was weighted to the comparative odds ratio associated with the risk factors. While this score is not proposed as a performance measure, we mention it here to show that a tool exists that can be used by hospitals to their bleeding rates and increase the safety of their PCI performance.

The C-statistic was used to describe the discrimination of the model and replicated in clinically important subgroups of interest, including patients STEMI, females, those aged >70 years, and patients with 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).

In 2017, NCDR updated its reporting to sites using hierarchical models and one component of the outcome (changing the threshold of a bleeding event in the absence of overt bleeding from a Hb drop of 3g/dl to 4g/dl to align with the bleeding definition in other registries, such as the ACTION AMI registry). The same predictor variables from the published model were used although the beta weights and intercepts were inappropriately updated. Furthermore, the performance characteristics of the model was confirmed. The extensive data provided in this submission, all run with the new model and bleeding definition, justifies the updated model.

**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?**

As described above, bivariate analyses were done to identify candidate variables that differed significantly between those with and without a clinically important bleeding event. Multivariable, hierarchical logistic regression analyses were then performed to retain those with a statistically significant association with bleeding (p<0.05 for each). Table 2 in Section 1.6 demonstrates the difference between those with and without bleeding events, based upon 2016 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.**

As noted in Section 1.8 above, social risk factors are not included in this clinically-focused measure.

**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*)

We developed the model in the 80% derivation set and tested its discrimination and calibration (using both the Hosmer-Lemeshow test and the slope of the predicted vs. observed risk). We then replicated this in 2 separate data sets; 20% of the original sample from 2/08-4/11 and in a completely unique set of data from 2016 (see above). Given secular trends in bleeding rates, with increasing use of radial approaches and bivalirudin leading to lower bleeding rates, we propose recalibrating the model with a new intercept (no change to the β-weights) each year, as was done for 2016 data.

*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 0.78 for the original model, which means that the probability that predicting the outcome is 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 in the table below:

**Table 4. Derivation and Validation C-Statistic**

|  | **N** | | **Full Model** | | **Risk Score** | |
| --- | --- | --- | --- | --- | --- | --- |
| **Group** | **Development**  **Sample** | **Validation**  **Sample** | **Development**  **Sample** | **Validation**  **Sample** | **Development**  **Sample** | **Validation**  **Sample** |
| Overall | 834,696 | 209,063 | 0.78 | 0.77 | 0.76 | 0.75 |
| STEMI | 133,649 | 33,311 | 0.71 | 0.71 | 0.70 | 0.70 |
| Women | 272,357 | 68,540 | 0.74 | 0.74 | 0.73 | 0.72 |
| Age >70 years | 275,089 | 69,015 | 0.76 | 0.76 | 0.74 | 0.74 |
| Diabetes | 299,402 | 75,003 | 0.78 | 0.78 | 0.76 | 0.76 |
| Excluding in-hospital CABG | 824,414 | 205,510 | 0.79 | 0.78 | 0.76 | 0.76 |
|  | | | | | | |

In the 2016 data, the c-statistic was 0.79, slightly higher than that observed in the original data. Comparable performance was observed across all socio-demographic and clinical subsets, as shown below:

|  |  |  |
| --- | --- | --- |
| Group | Sample Size | C-statistic |
| Overall | 717,510 | 0.790 |
| STEMI | 121,466 | 0.733 |
| Women | 223,562 | 0.742 |
| Age >70 yrs | 252,040 | 0.767 |
| Diabetes | 286,743 | 0.793 |
| Caucasian | 617,123 | 0.788 |
| Non-Caucasian | 100,387 | 0.796 |
| No-Insurance | 32,270 | 0.781 |

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

Before recalibrating the model to the 2016 data the slope of the calibration line was 1.059 (p<.001) indicating that the relationship between the independent variables in our model and the bleeding outcome slightly overpredicted bleeding in the lower risk patients, with a perfect slope being 1, and the intercept of the line was -0.1265 (p<.0001) indicating that the bleeding rate has decreased since the model was developed.

Due to the decreased bleeding rate from model development we recalibrated the model to the 2016 rates and obtained a slope and intercept of 1 and 0 respectively. See Figure 2.

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

**Figure 1. Calibration Curve Plot 2016**



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

For 2016 data, the risk stratification adequately segregated deciles of risk from <1% to >22% at the patient level. At the hospital level, we observed a broad range of unadjusted risk, which was partly mitigated after adjusting for patient characteristics. The unadjusted distribution of bleeding is shown under Figure 3.

S:\acc-ncdr\Analysis\2013\NQF\bleeding\2016\histo 2016 unadj.emf

**Figure 2. Unadjusted Distribution of Bleeding**

The bleeding rates adjusted for patient characteristics is shown under Figure 4.

S:\acc-ncdr\Analysis\2013\NQF\bleeding\2016\histo 2015 rsb.emf

**Figure 3. Adjusted Distribution of Bleeding**

After adjusting for patient characteristics, we observed a narrower and more normal distribution of bleeding outcomes.

The distribution of sites’ observed/expected ratios are shown under Figure 5.

S:\acc-ncdr\Analysis\2013\NQF\bleeding\2016\histo 2016 rratio.emf

**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 this model performs very well, accounting for patient characteristics present prior to the conduct of PCI and discriminating within important clinical subsets of patients. Moreover, there is substantial hospital variation before and after risk-adjustment. The distribution of institutional O/E ratios identifies some sites with excellent performance and others with rates of bleeding that are 80% or 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*)

We have provided extensive data about the model’s performance in much more recent data from which the model was originally developed, further supporting its robustness.

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

**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 bleeding rates across hospitals after adjusting for pre-procedural patient characteristics. Moreover, hospital performance on this measure is closely associated with risks for death, other complications and length of stay (Section 2b1.3).

**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 identifies the potential for improvement in comparison to others. Since all bleeding events are adverse outcomes, there are no absolute levels of bleeding risk that are significant as compared with others. To place the potential benefits of this model in context, it is helpful to compare the excess bleeds that could be avoided if the worst 25% of hospitals would have had the average bleeding rate of all hospitals. The average, adjusted bleeding rate was 3.2% and the upper quartile ranges from 3.9 to 13% bleeding rates. Given an average PCI volume of 410 cases/hospital, this suggests between 3 and 40 additional and potentially avoidable bleeding events per year among hospitals in the upper quartile as compared with the average hospital. This would be far larger if the worst performing hospitals were to achieve top-decile performance of a 1.8% bleeding rate. Clinically, these are a large number of excess events, particularly given that there are readily applied interventions, such as radial access, bivalirudin or the use of closure devices for femoral access, to mitigate bleeding.

**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*)  
 As noted above, there is minimal missing data due to the NCDR CathPCI submission requirements; missing data are imputed to include all cases in estimating performance.

**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*)

Missing data used to derive the bleeding model are minimal. The most frequent missing variables are pre-procedure hemoglobin (missing in 4.1%) and pre-procedure creatinine (missing rate in 3.6%). Both of these were imputed as medians according to patient sex and MI type. The other variables have missing rates under 0.5% and were imputed using median or most frequent category.

**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*)

Given the low rates of missing data, we do not believe that the observed performance is systematically biased. Our efforts to impute data had little effect on site’s performance on this measure.