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

**Measure Number** (*if previously endorsed*)**:** 2473

**Measure Title**: Hybrid hospital 30-day, all-cause, risk-standardized mortality rate (RSMR) following acute myocardial infarction (AMI)

**Date of Submission**: Click here to enter a date

**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 non-responders) 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*).

The dataset used for testing included Medicare Parts A and B claims, the Medicare Enrollment Database (EDB), data from the ACTION Registry®-GWTG™ (AR-G), census data, and electronically and manually abstracted electronic health record (EHR) data from several health systems.

During development of the measure, the registry data were used as a surrogate for data that will eventually come from electronic health records (EHRs). We subsequently established the feasibility of the EHR data elements in several health systems and EHR software environments. Additionally, census as well as claims data were used to assess socioeconomic factors (dual eligible obtained through enrollment data; Agency for Healthcare Research and Quality (AHRQ) socioeconomic status (SES) index score obtained through census data).

The dataset used varies by testing type; see Section 1.7 for details.

**1.3. What are the dates of the data used in testing**? The dates vary by testing type; see Section 1.7 for details.

**1.4. What levels of analysis** **were tested**? (*testing must be provided for all the levels specified and intended for measure implementation, e.g., individual clinician, hospital, health plan*)

|  |  |
| --- | --- |
| **Measure Specified to Measure Performance of:**  **(*must be consistent with levels entered in item S.20*)** | **Measure Tested at Level of:** |
| individual clinician | individual clinician |
| group/practice | group/practice |
| hospital/facility/agency | hospital/facility/agency |
| health plan | health plan |
| other: Click here to describe | other: American Community Survey |

**1.5. How many and which measured entities were included in the testing and analysis (by level of analysis and data source)**? (*identify the number and descriptive characteristics of measured entities included in the analysis (e.g., size, location, type); if a sample was used, describe how entities were selected for inclusion in the sample*)

The number of measured entities (hospitals) varies by testing type; see Section 1.7 for details.

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

The number of admissions/patients varies by testing type; see Section 1.7 for details.

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

The datasets, dates, number of measured hospitals and number of admissions used in each type of testing are as follows:

For reliability testing

The reliability of the model was tested by randomly selecting 50% of **Dataset 1** (development dataset) and developing a risk-adjusted model for this group. We then developed a second model for the remaining 50% of patients (validation sample) and compared the two. Thus, for reliability testing, we randomly split **Dataset 1** into two samples.

For measure development purposes only, we used two linked data sources to create **Dataset 1:** Medicare Administrative claims (Medicare Part A Inpatient and Outpatient claims data) merged with AR-G registry data. Registry data were used to obtain the clinical risk-adjustment variables that could in the future be extracted from EHRs.

**Dataset 1** (development dataset)

Dates of Data: January 1, 2009 – December 31, 2010

Number of Admissions: 54,736

Number of Measured Hospitals: 740

First half split sample (development sample)

Number of Admissions: 27,368

Number of Measured Hospitals: 280

Second half split sample (validation sample)

Number of Admissions: 27,367

Number of Measured Hospitals: 460

For validity testing (Section 2b2)

**Dataset 1** was used for measure validity testing

Three additional datasets were used to assess the feasibility and validity of several critical data elements.

**Dataset 2**: Data was provided from the administrative and EHR data warehouses of a large integrated health care delivery system that serves over 3.3 million members. All hospitals in this dataset used an integrated EHR system that runs Epic software.

* Number of admissions in dataset: 16,145
* Number of hospitals: 21
* Patient Descriptive Characteristics: mean age =58 with a standard deviation of 21 years; %female= 62.6

**Dataset 3**: Data were electronically extracted from one hospital that used Epic as their clinical EHR, and Siemens Invision A2K3 as their administrative EHR.

* Number of patients in EHR dataset: 23,624
* Number of patients in the data elements validation sub-sample of abstracted charts: 18,017
* Number of hospitals: 1

**Dataset 4**: Data were electronically extracted from one hospital that used Meditech as their clinical and administrative EHR.

* Number of patients in dataset: 1,853
* Number of patients in the data elements validation sub-sample of abstracted charts: 1,468
* Number of hospitals: 1

For testing of measure exclusions (Section 2b3)

**Dataset 1** (January 1, 2009 – December 31, 2010)

Number of Eligible Admissions: 217,723

Number of Eligible Measured Entities: 1,511

For testing of measure risk adjustment (Section 2b4)

**Dataset 1** (January 1, 2009 – December 31, 2010)

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

**Dataset 1 (**January 1, 2009 – December 31, 2010)

For testing of socioeconomic status (SES) factors and race in risk models (Section 2b4)

**Dataset 5** and **Dataset 6** (Section 2b4)

The impact of socioeconomic factors was not directly tested in the Hybrid AMI mortality measure due to lack of availability of EHR data from a nationally representative set of hospitals with patients who represent the full spectrum of socioeconomic status. Instead, we report results of testing done in the

Measure #0230, Hospital 30-day, all-cause, risk-standardized mortality rate (RSMR) following acute myocardial infarction (AMI) hospitalization for patients 18 and older.

**Dataset 5**: (2015 public reporting cohort): Medicare Part A Inpatient and Outpatient and Part B Outpatient claims Dates of Data: July 1, 2011 –June 30, 2014

Number of admissions: 497,550

Number of patients in sample A = 247,641

Number of patients in sample B = 249,909

Number of measured entities: 4,490

We examined disparities in performance according to the proportion of patients in each hospital who were dual eligible for both Medicare and Medicaid insurances. We also used the AHRQ SES index score to study the association between performance measures and SES.

**Dataset 6**: The American Community Survey (2008-2012)

We also used the Agency for Healthcare Research and Quality(AHRQ)SES index score derived from the American Community Survey (2008-2012) to study the association between performance measures and socioeconomic status.

Data Elements:

• Dual eligible status (i.e., enrolled in both Medicare and Medicaid) patient-level data are obtained from CMS enrollment data (Dataset 4)

• Validated AHRQ SES index score is a composite of 7 different variables found in the census data (the American Community Survey [2008-2012])

**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.

We selected social risk factors to analyze after reviewing the literature and examining available national data sources. Few patient-level social risk factors can be linked to Medicare data are available nationally. Dual-eligible status, e.g. enrolled in both Medicare and Medicaid [obtained from CMS claims enrollment data] is one of the only patient-level social risk variable available to examine directly.

We also considered neighborhood-level variables, linked by patient zip code level data that could serve in a risk model as a proxy for patient-level sociodemographic status (SDS). A range of census-collected social risk factor variables [collected annually as part of American Community Survey and aggregated over 5-years] including income and education, were available. We only linked the data at a 5-digit zip code level. Nine-digit zip code data may provide a more granular view of patient sociodemographic status, but this data is not available to us at the time of the analyses and we therefore cannot ascertain the incremental, if any, value of greater geographic discrimination for risk adjustment purposes.

Our conceptual model and the literature regarding how social risk factors may influence post-discharge mortality did not identify a single social risk factor as predominant in the pathway. There is a large body of literature linking various social risk factors to worse health status and higher mortality over a lifetime (Adler and Newman 2002, Mackenbach et al. 2000, Tonne et al. 2005, van Oeffelen et al. 2012). Income, education, and occupational level are the most commonly examined variables. However, literature directly examining how different social risk factors might influence the likelihood of mortality in older, insured, Medicare patients within 30 days of an admission for cardiovascular disease is much more limited. Assuming that the risk imparted based on zip code level data may reflect multiple different social risk variables, we chose to analyze a validated AHRQ composite index of socioeconomic status (SES), which has been used and tested among Medicare beneficiaries (Blum et al. 2014; Bonito et al. 2008). This index is a composite of 7 different variables found in the census data which may capture SES better than any single variable. The index variables include rates of unemployment, percent of person living below poverty, education level (percent below 12th grade education and percent with college education), crowding (average of more than one person per room), median household income, and median housing value. We identified patients as low SES if they lived in a neighborhood in the lowest quartile of this index.

Other variables can be found at a county or regional level and could represent the hospital’s community. We did not directly test any such variables because they are not as closely related to patients’ sociodemographic status given the wide scope of a county and seemed unlikely to be ideal for patient-level risk adjustment.

References

Adler NE, Newman K. Socioeconomic disparities in health: pathways and policies. *Health affairs (Project Hope).* 2002;21(2):60-76.

Blum AB, Egorova NN, Sosunov EA, et al. Impact of socioeconomic status measures on hospital profiling in New York City. Circulation. Cardiovascular quality and outcomes. May 2014;7(3):391-397.

Bonito A, Bann C, Eicheldinger C, Carpenter L. Creation of new race-ethnicity codes and socioeconomic status (SES) indicators for Medicare beneficiaries. Final Report, Sub-Task. 2008;2.

Mackenbach JP, Cavelaars AE, Kunst AE, Groenhof F. Socioeconomic inequalities in cardiovascular disease mortality; an international study. *European heart journal.* 2000;21(14):1141-1151.

Tonne C, Schwartz J, Mittleman M, Melly S, Suh H, Goldberg R. Long-term survival after acute myocardial infarction is lower in more deprived neighborhoods. *Circulation.* Jun 14 2005;111(23):3063-3070.

van Oeffelen AA, Agyemang C, Bots ML, et al. The relation between socioeconomic status and short-term mortality after acute myocardial infarction persists in the elderly: results from a nationwide study. *European journal of epidemiology.* Aug 2012;27(8):605-613.

**2a2. RELIABILITY TESTING**

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

**2a2.1. What level of reliability testing was conducted**? (*may be one or both levels*)  
 **Critical data elements used in the measure** (*e.g., inter-abstractor reliability; data element reliability must address ALL critical data elements*)  
 **Performance measure score** (e.g., *signal-to-noise analysis*)  
  
**2a2.2. For each level checked above, describe the method of reliability testing and what it tests** (*describe the steps―do not just name a method; what type of error does it test; what statistical analysis was used*)

Data Element Reliability: Electronic clinical data elements

See section 2b2 for validity testing of data elements.

Measure Score Reliability

The reliability of a measurement is the degree to which repeated measurements of the same entity agree with each other. For measures of hospital performance, the measured entity is the hospital, and reliability is the extent to which repeated measurements of the same hospital give similar results. In line with this thinking, our approach to assessing reliability is to consider the extent to which assessments of a hospital using different but randomly selected subsets of patients produces similar measures of hospital performance. That is, we take a "test-retest" approach in which hospital performance is measured once using a random subset of patients, then measured again using a second random subset exclusive of the first, and finally comparing the agreement between the two resulting performance measures across hospitals (Rousson et al., 2002).

For test-retest reliability, we combined index admissions from successive measurement periods into one dataset, randomly sampled half of patients within each hospital, calculated the measure for each hospital, and repeated the calculation using the second half. Thus, each hospital is measured twice, but each measurement is made using an entirely distinct set of patients. To the extent that the calculated measures of these two subsets agree, we have evidence that the measure is assessing an attribute of the hospital, not of the patients. As a metric of agreement we calculated the intra-class correlation coefficient (ICC), and assessed the values according to conventional standards (Landis and Koch, 1977; Shrout and Fleiss, 1979).

Specifically, we used **Dataset 1** split samples (development sample” and a “validation sample” and calculated the risk-standardized mortality rate (RSMR) for each hospital for each sample. As a metric of agreement we calculated the intra-class correlation coefficient (ICC) (Shrout and Fleiss, 1979), and assessed the values according to conventional standards (Landis and Koch, 1977). Specifically, we used Dataset 4 split sample and calculated the RSMR for each hospital for each sample. The agreement of the two RSRMs was quantified for hospitals using the intra-class correlation as defined by ICC [2,1] by Shrout and Fleiss (1979).

Using two independent samples provides a stringent estimate of the measure’s reliability, compared with using two random but potentially overlapping samples which would exaggerate the agreement.

Moreover, because our final measure is derived using hierarchical logistic regression, and a known property of hierarchical logistic regression models is that smaller volume hospitals contribute less ‘signal´, a split sample using a single measurement period would introduce extra noise. This leads to an underestimate in the actual test-retest reliability that would be achieved if the measure were reported using the full measurement period, as evidenced by the Spearman Brown prophecy formula (Spearman 1910, Brown 1910). We use this to estimate the reliability of the measure if the whole cohort were used, based on an estimate from half the cohort.

Test-retest reliability is considered the lower bound of any reliability estimate (Yu, Mehrotra, and Adam, 2013). While it is the most relevant metric from the perspective of measure reliability, it is also meaningful to consider the separate notion of “unit” reliability, that is, the reliability with which individual units (here, hospitals) are measured. Therefore, we also use the approach used by Adams and colleagues to calculate reliability for this measure (2010). Because this metric has been reported for other measures in other contexts (see e.g., Adams et al 2010), and to provide an additional, complementary metric, we also report this average unit reliability.

References

AdamsJ, Mehrota, A, Thoman J, McGlynn, E. (2010). Physician cost profiling – reliability and risk of misclassification. NEJM, 362(11): 1014-1021.

Brown, W. (1910). Some experimental results in the correlation of mental abilities. British Journal of Psychology, 3, 296-322.

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

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

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

Spearman, Charles, C. (1910). Correlation calculated from faulty data. British Journal of Psychology, 3,271–295.

Yu, H, Mehrota, A, Adams J. (2013). Reliability of utilization measures for primary care physician profiling. Healthcare, 1, 22-29.

**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*)

Reliability of Measure Score

There were 54,735 admissions in the measure cohort, with 27,368 in one randomly selected sample (development sample) and 27,367 in the other sample (validation sample). The agreement between the two RSMRs for each hospital was 0.42, which according to the conventional interpretation is “moderate” (Landis & Koch, 1977).

Please note that the above reliability represents the lower bound of any reliability estimate of this measure. Using the approach by Adams et al (2010), we found that among the 375 hospitals with 25 and more cases in the combined two years data of 2009 and 2010, both the median and mean reliability are 0.543. This is considered to be moderate (Landis & Koch, 1977).

Reference:

AdamsJ, Mehrota, A, Thoman J, McGlynn, E. (2010). Physician cost profiling – reliability and risk of misclassification. NEJM, 362(11): 1014-1021.

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

**2a2.4 What is your interpretation of the results in terms of demonstrating reliability**? (i*.e., what do the results mean and what are the norms for the test conducted?*)

Reliability of Measure Score

For the hospital event rate based on the patient binomial outcomes like readmission (Yes/No), an ICC value of 0-0.2 indicates poor agreement; 0.3-0.4 indicates fair agreement; 0.5-0.6 indicates moderate agreement; 0.7-0.8 indicates strong agreement; and >0.8 indicates almost perfect agreement. The ICC of 0.42 demonstrates fair agreement across samples using a conservative approach to assessment for the measure score.

The ICC[2,1] is a conservative measure of test-retest reliability because it assumes that the multiple measurements are drawn from a larger sample of tests, and that the measured providers are drawn from a larger sample of providers. Given, the conservative nature of the ICC[2,1] and the complex constructs of the measure itself, a lower reliability score is expected.

Guidelines for the interpretation of the ICC[2,1] statistic are limited. Landis & Koch (Landis, Koch 1977) created a convention to assess the reliability but stated “In order to maintain consistent nomenclature when describing the relative strength of agreement associated with kappa statistics, the following labels will be assigned to the corresponding ranges of kappa… Although these divisions are clearly arbitrary, they do provide useful “benchmarks” for the discussion of the specific example in Table 1”.

In other words, ‘acceptability’ depends on context. For example, if we were measuring adolescent weight twice with the same scale, and assessing whether the weights were above a certain threshold, we would expect the two measurements to agree almost exactly (ICC[2,1] ~ 1); otherwise, we would discard the scale. At the other extreme, if we were measuring a latent personality trait such as a personality disorder, we would expect a much lower level of agreement. In fact, Nestadt et al assessed ICCs for several standard tools for assessing personality disorder and found test-retest reliabilities in the range of 0.06-0.27 (Nestadt 2012). Notably, Nestadt et al conclude that these tools “may still be useful for identifying [personality disorder] constructs.”

The current context is measuring provider quality, or specifically provider propensity to provide appropriate care as measured by subsequent outcomes. Cruz et al report reliabilities for collecting risk factor information from patients presenting to an emergency department with potential acute coronary syndrome (ACS) [Cruz et al]. Each patient was queried twice, once by a clinician and once by research assistant, and the reliabilities for a range of risk factors were calculated; these ranged from 0.28 (associated symptoms) to 0.69 (cardiac risk factors), with all other factors in the 0.30-0.56 range. Hand et al report test-retest reliabilities for bedside clinical assessment of suspected stroke [Hand et al]. Pairs of observers independently assessed suspected stroke patients; findings were recorded on a standard form to promote consistency. The reliabilities were calculated for the full range of diagnostic factors: for vascular factors reliabilities ranged from 0.47-0.69 with only four of eight above 0.6; for history, they ranged from 0.37-0.65 with only five of 12 above 0.6; other categories were similar (though reliability=1 for whether the patients were conscious).

Given the limited resources available, the arbitrary nature of divisions, and the current literature, we feel that there is sufficient reliability in the measure score.

**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)*

Validity of EHR Data Elements

Several critical clinical data elements used in the measure’s risk models were derived from patients’ electronic medical records. When this measure is implemented, CMS intends to obtain these critical data elements from hospital EHRs and merge the data with claims data to calculate and report measure results. We tested the validity of electronic extraction of these critical data elements as part of a more comprehensive evaluation of a larger set of core clinical data elements (CCDEs). The CCDE are a set of 21 EHR data elements that are captured on most adults (plus Troponin, which is a condition-specific CCDE for patients with acute myocardial infarction) admitted to acute care hospitals, are easily extracted from EHRs, and can be used to risk adjust hospital outcome measures for a variety of conditions and procedures. All of the critical data elements used in the hybrid AMI mortality measure are included in the CCDE. Testing of the CCDE involved three phases: 1) identification of potentially feasible clinical data through qualitative assessment, 2) empirical feasibility testing of several clinical data elements electronically extracted from two large multi-facility health systems, and 3) validity testing of the CCDE at two additional health systems.

**Phase 1: Identification of potentially feasible clinical data through qualitative assessment**

To identify the CCDEs for risk adjustment of hospital outcome measures for adult patients, we first conducted a qualitative assessment of the reliable capture, accuracy, and extractability of categories and subcategories of clinical data as defined by the Quality Data Model (QDM) (e.g., vital signs, laboratory test results). We established a set of criteria to assess the consistency of data capture, relevance to hospital quality measures, and extractability from health records.

Data Capture Criteria:

Obtained consistently under current practice. Routinely collected for patients admitted to the hospital under current clinical practice and EHR workflows.

Captured with a standard definition. Consistent conceptual understanding, method of collection, and units of measurement.

Entered in a structured field. Captured in numerical, pseudo-numerical, or list format.

Data Extraction Criteria:

Encoded consistently. Can be linked to a standard and uniform coding structure such as ICD-9 or Systematized Nomenclature of Medicine Clinical Terms (SNOMED-CT).

Extractable from the EHR. Can be readily and consistently identified and exported from current EHR databases.

Exported with metadata. Additional information such as time stamps and reference values that are needed for interpretation are consistently available.

These criteria are aligned with those established in the NQF’s Hybrid Feasibility Assessment Report as well as the NQF feasibility criteria (see included Data Element Feasibility Scorecard). The NQF report emphasized four key aspects of feasibility. First, data should be structured or easily converted to a structured and interpretable format. Second, data should be accurate. Third, data should be easily associated with a standard set of codes to ensure consistent extraction across EHR environments. Finally, data should not require changes to current clinical practice or workflows.

We then convened a Technical Expert Panel (TEP) to apply these criteria to categories and subcategories (data types) of clinical data based on the Quality Data Model (QDM). We asked TEP members to consider only the context of adult hospitalized patients when making their assessments. Data categories and subcategories were rated on each feasibility criterion independently by TEP members. The ratings were tallied and TEP members met to discuss and resolve areas of disagreement. Through this process the TEP identified a list of data subcategories that were potentially feasible for use in hospital outcome measures. The CCDE were derived from only those subcategories for which the TEP reached consensus agreement on feasibility.

**Phase 2: Empirical feasibility testing using a large multi-site database (Dataset 2)**

We next directly examined the feasibility of clinical data elements from the subcategories identified by the TEP as feasible (for all adult inpatient admissions). We used a three-year dataset that contained merged inpatient claims with clinical data elements derived from patients’ EHRs from a single health system (**Dataset 2**). These data were extracted from an Epic EHR system. The merged data were provided for all patients discharged from any of the 21 acute care hospitals within the health system from January 1, 2010 through December 31, 2012. We examined all admissions to ensure they were captured in a numerical field, the consistency and timing of capture, and the accuracy of the data elements. We examined the data elements across conditions, hospitals, and point of hospital entry. We tested several data elements that met the feasibility criteria in models predicting 30-day mortality following admission for several common medical conditions. The complete list of 21 (plus Troponin) CCDE were derived from these analyses, including the subset of five CCDE that are used in the hybrid AMI mortality measure.

Additionally, we assessed the rate and timing of capture of the data elements in **Dataset 1** and **Dataset 2**.

**Phase 3: Validity testing of the CCDE at two hospital sites (including critical data elements for the hybrid AMI mortality measure)**

In Phase 3, we developed electronic specifications (e-specifications) using the Measure Authoring Tool (MAT), and analyzed extracted data from EHRs. We assessed the ability of hospitals to use the e-specifications to query and electronically extract CCDEs from the EHR, for all adult inpatient admissions occurring over the course of one year. Validity testing assessed the accuracy of the electronically extracted CCDEs compared to the same CCDEs gathered through manual abstraction (from the EHR) in a subset of 23,624 charts identified in the data query in **Dataset 3**, and 1,853 charts identified in the data query in **Dataset 4**.

*Chart Abstraction*: We calculated the number of admissions that needed to be randomly sampled from the EHR dataset and manually abstracted to yield a statistical margin of error (MOE) of 5% and a confidence level of 95% for the match rates between the two data sources. Sites then used an Access-based manual abstraction tool provided (along with training) to manually abstract the CCDEs from the random samples of the medical records identified through the EHR data query. The manual chart abstraction data is considered the “gold standard” for the purpose of this analysis.

*Validity Testing*: We conducted validity testing on the critical EHR data elements in the Hybrid AMI mortality measure. For each continuous data element, we were only interested in the case where the electronic abstraction value exactly matched the manual abstraction value. We therefore only calculated the raw agreement rate between data from electronic and manual chart abstraction. For simple data values, we believe taking this approach, as compared to reporting statistical tests of accuracy, better reflects the concept of matching exact data values rather than calculated measure results. Therefore, we do not report statistical testing of the accuracy of the EHR derived data value as compared with the abstracted value. Instead, we counted only exact matches in the data value as well as the time and date stamp associated with that value when we calculated the match rate. The 95% confidence level was established based on the sample size and reflects the exact match rate using these criteria.

Validation of the Measure Score Compared with Other Risk Models and Registry Data

We compared the hospital-level results from this hybrid AMI mortality measure to the results from the harmonized claims-only measure #0230*, Hospital 30-day, all-cause, risk-standardized mortality rate (RSMR) following acute myocardial infarction (AMI) hospitalization for patients 18 and older*. Both models use inpatient administrative and claims data to derive the cohort, and to assess the outcome.

Measure validity was tested through comparison of this Hybrid risk adjustment model with claims-only risk-adjustment model, and through use of established measure development guidelines.

For the derivation of both risk models, we used **Dataset 1** (development sample). Both the Hybrid and claims-only risk models used the same inclusion/exclusion criteria and a risk-adjustment (statistical modeling) strategy and only differed with respect to the risk variables used. We compared the model discrimination and the correlation in hospital performance results for the two models.

Validity Indicated by Established Measure Development Guidelines

We developed this measure in consultation with national guidelines for publicly reported outcomes measures, with outside experts, and with the public. The measure is consistent with the technical approach to outcomes measurement set forth in NQF guidance for outcomes measures (National Quality Forum, 2010), CMS Measure Management System (MMS) guidance, and the guidance articulated in the American Heart Association scientific statement, “Standards for Statistical Models Used for Public Reporting of Health Outcomes” (Krumholz, Brindis, et al., 2006).

Validity as Assessed by External Groups

Throughout measure development, we obtained expert and stakeholder input via regular discussions with an advisory working group and a 30-day public comment period in order to increase transparency and to gain broader input into the measure.

The working group was assembled, and regular meetings were held throughout the development phase. The working group was tailored for development of this measure and consisted of clinicians (cardiologists) and other professionals with expertise in biostatistics, measure methodology, and quality improvement. The working group meetings addressed key issues related to measure development, including the deliberation and finalization of key decisions (e.g., defining the measure cohort and outcome) to ensure the measure is meaningful, useful, and well-designed. The working group provided a forum for focused expert review and discussion of technical issues during measure development.

Following completion of the preliminary model, we solicited public comment on the measure through the CMS site link: https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/MMS/Public-Comments.html. The public comments were then posted publicly for 30 days. The resulting input was taken into consideration during the final stages of measure development and contributed to minor modifications to the measure.

References:

Bratzler DW, Normand SL, Wang Y, et al. An administrative claims model for profiling hospital 30-day mortality rates for pneumonia patients. PLoS One 2011;6(4):e17401.

Keenan PS, Normand SL, Lin Z, et al. An administrative claims measure suitable for profiling hospital performance on the basis of 30-day all-cause readmission rates among patients with heart failure. Circulation 2008;1(1):29-37.

Krumholz HM, Brindis RG,Brush JE, et al. Standards for Statistical Models Used for Public Reporting of Health Outcomes: An American Heart Association Scientific Statement From the Quality of Care and Outcomes Research Interdisciplinary Writing Group: Cosponsored by the Council on Epidemiology and Prevention and the Stroke Council Endorsed by the American College of Cardiology Foundation. Circulation. January 24, 2006 2006;113(3):456-462.

Krumholz HM, Wang Y, Mattera JA, et al. An administrative claims model suitable for profiling hospital performance based on 30-day mortality rates among patients with an acute myocardial infarction. Circulation 2006;113(13):1683-92.

Krumholz HM, Wang Y, Mattera JA, et al. An administrative claims model suitable for profiling hospital performance based on 30-day mortality rates among patients with heart failure. Circulation 2006;113:1693-1701.

National Quality Forum. National voluntary consensus standards for patient outcomes, first report for phases 1 and 2: A consensus report http://www.qualityforum.org/projects/Patient\_Outcome\_Measures\_Phases1-2.aspx. Accessed August 19, 2010.

Shahian DM, He X, O’Brien S, et al. Development of a Clinical Registry-Based 30-Day Readmission Measure for Coronary Artery Bypass Grafting Surgery. Circulation 2014; DOI: 0.1161/CIRCULATIONAHA.113.007541. Published online before print June 10, 2014

Suter L, Wang C, Araas M, et al. Hospital-Level 30-Day All-Cause Unplanned Readmission Following Coronary Artery Bypass Graft Surgery (CABG): Updated Measure Methodology Report. 2014; http://www.qualitynet.org/dcs/BlobServer?blobkey=id&blobnocache=true&blobwhere=1228890352615&blobheader=multipart%2Foctet-stream&blobheadername1=Content-Disposition&blobheadervalue1=attachment%3Bfilename%3DRdmsn\_CABG\_MeasMethd\_Rpt\_060314.pdf&blobcol=urldata&blobtable=MungoBlobs. Accessed November 4, 2015.

ICD-9 to ICD-10 Conversion

Statement of Intent

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

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

[ ] The intent of the measure has changed.

Process of Conversion

This cohort (inclusions and exclusions) for this hybrid measure is defined using ICD-CM codes. This hybrid measure cohort is fully harmonized with measure #0230, Hospital 30-day, all-cause, risk-standardized mortality rate (RSMR) following acute myocardial infarction (AMI) hospitalization for patients 18 and older which is publically reported in the Inpatient Quality Reporting Program.

We re-specified the measure cohort to accommodate the implementation of ICD-10 coding. Specifically:

• We expanded the cohort definition to include ICD-10 codes for use with discharges on or after October 1, 2015. (Previously-specified ICD-9 codes continue to be used for discharges before October 1, 2015.)

The goal of this re-specification was to maintain the intent and validity of the measure.. In developing the ICD-10 code lists that define the cohort for the measure, we created cohort crosswalks using the General Equivalence Mappings (GEMs), a tool created by CMS and the Centers for Disease Control and Prevention (CDC) to assist with the conversion of ICD-9 codes to ICD-10 codes (Part of The ICD-10 Transition Process). To validate the cohort crosswalks, we compared the cohort size using ICD-10 codes in a set of claims submitted between October 2015 and March 2016 with the cohort size using previously-defined ICD-9 codes in aset of claims submitted between October 2014 and March 2015. We conducted clinical review of the results of this analysis to further refine the set of codes appropriate for cohort definition.

CD-9 and ICD-10 codes are attached in the Data Dictionary.

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

Data Element Validity

We assessed data element validity of the hybrid AMI mortality measure using the percent agreement between findings of electronic extraction and manual abstraction in the EHR systems at three hospitals as follows:

**Phase 1: TEP Survey Results**

The TEP identified seven subcategories of EHR data that they considered feasible for adult hospitalized patients. They were: Encounter Performed, Patient Characteristics including birth date and sex, Physical Examination Findings for vital signs only, Diagnostic Study Order, Diagnostic Study Performed, Medication Discharge, and Laboratory Test Result. We limited the CCDE to data elements to only four categories: Encounter Performed, Patient Characteristics, Physical Examination Findings for vital signs only, and Laboratory Test Results, which are unlikely to be reflective of care quality and therefore are thought to be both feasible to extract and appropriate for risk adjustment.

**Phase 2: Feasibility Testing Results**

**Datasets 2, 3, and 4**: The Table below shows the consistency of data capture of the critical data elements included in the Hybrid AMI mortality measure for all adult hospitalized patients in Dataset 2 where initial data element feasibility was tested, as well as in **two health systems** that extracted the data elements using the MAT output in two different EHR environments (EPIC and Meditech).

**Table. Percent Captured per Data Element per Hospital (Datasets 2, 3 and 4)**

| Data Element/ CCDE | % Captured  Dataset 1 | % Captured  Dataset 2 | % Captured  Dataset 3 | % Captured  Dataset 4 |
| --- | --- | --- | --- | --- |
| Heart Rate (BPM) | 99.86 | 97.7 - 97.9 | 84.73 | 98.97 |
| Systolic Blood Pressure (mmHG) | 99.86 | 97.6 – 97.8 | 84.61 | 99.02 |
| Creatinine | 99.51 | 95.2 – 95.3 | 88.90 | 92.00 |
| Troponin | 98.29 |  | 94.1 | 83.3 |

See the feasibility scorecard for additional assessment of data element feasibility.

**Phase 3: Further Feasibility and Validity Testing Results**

Chart abstraction for validity testing was done in **Dataset 3** and **Dataset 4**. The Tables below demonstrate the agreement in data values, time and date stamps between electronically extracted and manually abstraction data elements from the two health systems **(Dataset 3 and Dataset 4).**

**Table. Percent Agreement and Confidence Intervals: Comparison of EHR-Extracted and Manually Abstracted CCDE (Dataset 3, N=91)**

| CCDE | % agreement between electronic and manual data sets (#) | 95 percent confidence interval for agreement | Total # of admissions successfully compared between data sets | % present in electronic extraction, missing in manual abstraction (#) | % present in manual abstraction, missing in electronic extraction (#) | % missing in both electronic extraction and manual abstraction (#) |
| --- | --- | --- | --- | --- | --- | --- |
| Physical Exam/Vital Signs | | | | | | |
| Heart rate (BPM) | 90.79 (69) | 0.84 - 0.97 | 76 | 1.10 (1) | 1.10 (1) | 14.29 (13) |
| Systolic blood pressure (mmHG) | 89.47 (68) | 0.82 - 0.97 | 76 | 1.10 (1) | 1.10 (1) | 14.29 (13) |
| Laboratory Results | | | | | | |
| Creatinine (mg/dL) | 93.15 (68) | 0.87 - 0.99 | 73 | 0.00 (0) | 0.00 (0) | 19.78 (18) |
| Troponin\* (ng/mL) | 95.24 (20) | 0.85 - 1.05 | 21 | 4.40 (4) | 0.00 (0) | 72.53 (66) |

\* Troponin was only compared in 21 admissions because we exclusively tested agreement in patients with a principal discharge diagnosis of AMI

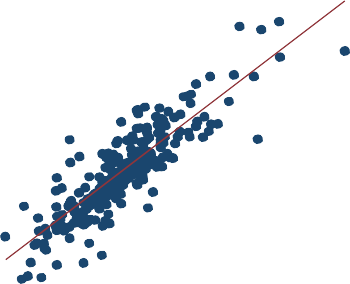
**Table.** **Percent Agreement and Confidence Intervals: Comparison of EHR-Extracted and Manually Abstracted CCDE (Dataset 4, N=92)**

| CCDE | % agreement between electronic and manual data sets (#) | 95 percent confidence interval for agreement | Total # of admissions successfully compared between data sets | % present in electronic extraction, missing in manual abstraction (#) | % present in manual abstraction, missing in electronic extraction (#) | % missing in both electronic extraction and manual abstraction (#) |
| --- | --- | --- | --- | --- | --- | --- |
| Physical Exam/Vital Signs | | | | | | |
| Heart rate (BPM) | 91.21 (83) | 0.85 - 0.97 | 91 | 0.00 (0) | 0.00 (0) | 1.09 (1) |
| Systolic blood pressure (mmHG) | 92.31 (84) | 0.87 - 0.98 | 91 | 0.00 (0) | 0.00 (0) | 1.09 (1) |
| Laboratory Results | | | | | | |
| Creatinine (mg/dL) | 86.05 (74) | 0.79 - 0.94 | 86 | 0.00 (0) | 5.43 (5) | 1.09 (1) |
| Troponin (ng/mL) | 89.19 (33) | 0.79 - 1.00 | 37 | 5.43 (5) | 2.17 (2) | 52.17 (48) |

Validation of the Measure Score Compared with Claims-Only Risk Model **(Dataset 1, development sample)**

We calculated the correlation of the RSMR from our final model with that of the previously validated, publicly reported claims-based AMI mortality measure, using data from 2009.

Figure 1. Correlation of the AMI mortality hybrid RSMRs and RSMRs based on the previously developed, publicly reported claims-based AMI mortality measure (hospital volume-weighted Pearson correlation coefficient=0.86)



.09

.1

.11

.12

.13

.14

RSMR -- Final Model

**2b1.4. What is your interpretation of the results in terms of demonstrating validity**? (i*.e., what do the results mean and what are the norms for the test conducted?*)

Data Element Validity

Data element validity testing of the fully eSpecified AMI mortality hybrid supported the overall validity of nearly all of the data elements included in the hybrid. All data elements for cohort identification and risk adjustment were consistently found for all patients and were both extractable and accurate. The critical data elements were demonstrated to be feasible through consensus of the TEP and direct examination of EHR data establishing consistent capture of the CCDE among adult hospitalized patients. In addition, we established the validity of electronic extraction of the CCDE demonstrated by the high match rate when comparing EHR extracted and manual medical record abstracted CCDE values.

Performance measure score: Empirical validity testing

The correlation coefficient of 0.86 demonstrates excellent correlation between the Hybrid and the claims-based AMI mortality measure. Measure validity was also ensured through the processes employed during development, including regular expert and clinical input.

**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*)

All exclusions were determined by careful clinical review and have been made based on clinically relevant decisions and to ensure accurate calculation of the measure. To ascertain impact of exclusions on the cohort, we examined overall frequencies and proportions of the total cohort excluded for each exclusion criterion (**Dataset 1**). These exclusions are consistent with similar NQF-endorsed outcome measures. Rationales for the exclusions are detailed in the Measure Submission Form (see section on Denominator Exclusions).

**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 **Dataset 1** (prior to exclusions being applied):

**Exclusion N %**

1. Discharged against medical advice (AMA) 53 0.24%

2. Transferred in from another short-term acute 615 2.80%

care institution

3. Unknown death records with missing vital status) 0 0.0%

in Medicare Enrollment Database

4. Unreliable data 1 0.00%

5. Multiple AMI admissions in 2009 431 2.00%

**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*)

**Exclusion 1** (patients who are discharged AMA) accounts for 0.24% of all index admissions excluded from the initial index cohort. This exclusion is needed for acceptability of the measure to hospitals, who do not have the opportunity to deliver full care and prepare the patient for discharge.

**Exclusion 2** (patients transferred in from another federal hospital) accounts for 0.24% of all index procedures excluded from the initial index cohort. This exclusion is intended to remove admissions from the cohort for patients transferred to federal hospitals. It is necessary for valid calculation of the measure. Very few patients are affected by this exclusion.

**Exclusion 3** (unknown death records) and **Exclusion 4** (unreliable data) account for 0% and 1% of all index admissions excluded from the initial index cohort. These exclusions affect very few patients and are need for valid calculation of the measure.

**Exclusion 5** (multiple AMI admissions) accounts for and 2% of all index procedures excluded from the initial index cohort. This exclusion is needed to ensure that episodes are independent for statistical purposes

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

**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** 5 **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.**

The risk model specification and methodology are described in Section 2b3.3a and the attached data dictionary.

**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?

Our approach to risk adjustment was tailored to and appropriate for a publicly reported outcome measure, as articulated in the American Heart Association (AHA) Scientific Statement, “Standards for Statistical Models Used for Public Reporting of Health Outcomes” (Krumholz et al., 2006).

The measure employs a hierarchical logistic regression model (a form of hierarchical generalized linear model [HGLM]) to create a hospital-level 30-day RSMR. This approach to modeling appropriately accounts for the structure of the data (patients clustered within hospitals), the underlying risk due to patients’ comorbidities, and sample size at a given hospital when estimating hospital mortality rates. In brief, the approach simultaneously models two levels (patient and hospital) to account for the variance in patient outcomes within and between hospitals (Normand and Shahian et al., 2007). At the patient level, each model adjusts the log-odds of mortality within 30-days of admission for age, selected clinical covariates and a hospital-specific intercept. The second level models the hospital-specific intercepts as arising from a normal distribution. The hospital intercept, or hospital-specific effect, represents the hospital contribution to the risk of mortality, after accounting for patient risk and sample size, and can be inferred as a measure of quality. The hospital-specific intercepts are given a distribution in order to account for the clustering (non-independence) of patients within the same hospital. If there were no differences among hospitals, then after adjusting for patient risk, the hospital intercepts should be identical across all hospitals.

Clinical Factors

To create a model with increased usability while retaining excellent model performance, we tested the performance of the model without those variables considered to be questionably feasible. Based on the results of that testing, the final parsimonious risk-adjustment model consisted of five variables that were clinically relevant and deemed to be hybrid-feasible.

During model development using **Dataset 1**, we performed a bootstrap simulation with 1,000 iterations by allowing patients to be selected repeatedly. In each iteration, a bootstrap data sample was constructed and a logistic regression model with stepwise selection (entry variables with p<0.05; retained variables with p<0.01) was performed over all the candidate variables.

The working group reviewed the results of the bootstrap simulation and decided to retain all risk-adjustment variables above a 90% cutoff (i.e., the variables were selected as significant at p<0.05 in 90% of the iterations), which was thought to demonstrate a consistently strong association with mortality. After running the bootstrap simulation on 22 candidate variables, the preliminary risk-adjustment model consisted of nine variables. Four of these had questionable feasibility (see 2b4.3) and were excluded from the final model.

The final risk model includes:

Age (years)

Heart Rate: HR<70 (10 bpm)

Heart Rate: HR>=70 (10 bpm)

Systolic Blood Pressure (10 mm Hg)

Troponin Ratio\* (ng/mL) (per 10 units) which is Initial troponin value (ng/mL) divided by the Troponin upper range limit (ng/mL)

Creatinine (mg/dL)

**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) We describe analysis done using data from measure #0230**

**Hospital 30-day, all-cause, risk-standardized mortality rate (RSMR) following acute myocardial infarction (AMI) hospitalization for patients 18 and older.**

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

**Acute Myocardial Infarction (AMI) Mortality Final Model: Hierarchical Logistic Regression Model Results (N=20,540 patients)**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variable | Estimate | SE | P value | O.R. | 95% CI |
| Age (years) | 0.063 | 0.003 | 0.000 | 1.07 | 1.06, 1.07 |
| Heart Rate: HR<70 (10 bpm) | -0.050 | 0.040 | 0.214 | 0.95 | 0.88, 1.03 |
| Heart Rate: HR>=70 (10 bpm) | 0.149 | 0.012 | 0.000 | 1.16 | 1.13, 1.19 |
| Systolic Blood Pressure (10 mm Hg) | -0.249 | 0.010 | 0.000 | 0.78 | 0.76, 0.80 |
| Troponin Ratio\* (ng/mL) (per 10 units) | 0.121 | 0.011 | 0.000 | 1.13 | 1.11, 1.15 |
| Creatinine (mg/dL) | 0.670 | 0.036 | 0.000 | 1.95 | 1.82, 2.10 |
|  |  |  |  |  |  |

280 hospitals with between-hospital variance=0.0248, standard error=0.0143

\*Troponin Ratio=Initial troponin value (ng/mL) / Troponin upper range limit (ng/mL)

Note: these results were calculated using the registry model development dataset with data from the 2009 calendar year only, and were validated in the 2010 calendar year registry 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.**

Because the hybrid AMI measure was developed using patients’ data from a subset of all of the nation’s hospitals, analysis of the potential impact of social risk variables could be distorted and provide a poor representation of the results if all of the nation’s hospitals were included. Because of this potential lack of representativeness and due to the high degree of correlation of the results of this measure with the results of the claims-based AMI measure, we have presented the results of analyses using claims data from the Hospital 30-day, all-cause, risk-standardized mortality rate (RSMR) following acute myocardial infarction (AMI) hospitalization for patients 18 and older.

Variation in prevalence of the factor across measured entities  
The prevalence of social risk factors in the AMI cohort varies substantially across hospitals. The median percent of dual eligible patients is 10.8% (interquartile range [IQR] 6.9%-16.8%). The median frequency of low SES AHRQ indicator patients is 16.4% (IQR 4.1%-40.3%).

Empirical association with the outcome (bivariate)

The patient-level observed AMI unadjusted mortality rate for dual-eligible patients was somewhat higher, at 16.1% compared with 14.0% for all other patients. The mortality rate for patients in the lowest SES quartile by AHRQ Index was slightly higher at 14.4% compared with 13.9% for patients in the highest SES quartile.

Incremental effect of SDS variables in a multivariable model

We then examined the strength and significance of the SDS variables in the context of a multivariable model. Each of the variables remained significantly associated in the multivariable model.

For dual eligibility and the AHRQ SES indicator, the variable is associated with higher risk of modest strength. Odds ratios are on the order of 1.12 for dual eligibility and 1.09 for AHRQ SES. This is similar to the odds ratio for comorbidities such as COPD and substantially lower than the risk associated with comorbidities such as metastatic cancer. In all cases, the c-statistic for the AMI patient-level multivariate model with the SDS variable in the model is essentially unchanged from that without the variable.

To further understand the relative importance of these risk-factors in the measure we compared hospital performance with and without the addition of each SDS variable. We found that the addition of any of these variables into the model has little to no effect on hospital performance. The mean absolute change in hospitals’ RSMRs when adding a dual eligibility indicator is -0.00039% with a correlation coefficient between RSMRs for each hospital with and without dual eligibility added of 0.9996. The mean absolute change in hospitals’ RSMRs when adding a low SES AHRQ indicator is -0.00205% with a correlation coefficient between RSMRs for each hospital with and without low SES added of 0.9982.

Overall, we found that among the SDS variables that could be feasibly incorporated into this model, 1) the relationship with mortality is small. We also found that the impact of adding any of these indicators is very small to negligible on model performance and hospital profiling.

Given these findings in the AMI Mortality claims-based measure and complex pathways that could explain any relationship between SDS and mortality, which do not all support risk-adjustment, we did not incorporate SDS variables into the 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*)  
*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)

During measure development, we computed the following summary statistics for assessing model performance (Harrell and Shih, 2001) for Dataset 1 (development sample & validation sample):

(1) Area under the receiver operating characteristic (ROC) curve

(2) Adjusted R-squared

(3) Predictive ability

(4) Calibration

We tested the performance of the model developed in a randomly selected 50% sample of **Dataset 1** (development sample) by comparing results with those from the validation sample (dataset).

References:

F.E. Harrell and Y.C.T. Shih. Using full probability models to compute probabilities of actual interest to decision makers. Int. J. Technol. Assess. Health Care 17 (2001), pp. 17–26.

**2b3.6. Statistical Risk Model Discrimination Statistics** (*e.g., c-statistic, R-squared*)**:**

Model performance was similar in the development and validation datasets, with strong model discrimination and fit. Predictive ability was also similar across datasets. The c-statistic (area under the ROC curve) was 0.78 for both datasets.

The adjusted R-squared was 0.204 for the development sample (data from 2009 and 0.194 for the validation sample (data from 2010)

Predictive Ability at the lowest decile % and highest decile % was 0.012 and 0.375 for the development sample, and 0.012 and 0.374 for the validation sample.

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

Calibration was

• -0.000, 1.000 for the development sample and

• -0.013, 0.979 for the validation sample

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

The risk decile plot is a graphical depiction of the observed mortality in the deciles of the predicted mortality to measure predictive ability. Below, we present the risk decile plot showing the distributions for the development dataset (**Dataset 1**). The plot for the validation dataset was similar.

Table. Model Performance: Risk decile plots

|  |  |  |
| --- | --- | --- |
| **Indices** | **2009 Development Sample** | **2010 Validation Sample** |
| Number of Admissions | 20,540 | 34,196 |
| Predictive Ability by Decile (%) | | |
| 1 | 1.2 | 1.2 |
| 2 | 2.7 | 2.4 |
| 3 | 2.9 | 4.0 |
| 4 | 4.7 | 4.9 |
| 5 | 4.7 | 5.5 |
| 6 | 7.5 | 8.1 |
| 7 | 10.9 | 9.8 |
| 8 | 13.3 | 14.4 |
| 9 | 22.5 | 22.1 |
| 10 | 37.5 | 37.4 |

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

N/A

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

***Discrimination Statistics***

The c-statistic of 0.78 indicates excellent model discrimination.

***Calibration Statistics***

*Over-fitting (Calibration γ0, γ1)*

The calibration value of close to 0 at one end and close to 1 to the other end indicates good calibration of the model.

***Risk Decile Plots***

The risk decile plot shows excellent discrimination of the model and good predictive ability.

***Overall Interpretation***

Interpreted together, our diagnostic results demonstrate the hybrid risk-adjustment model adequately controls for differences in patient characteristics (case mix).

**2b3.11.** **Optional Additional Testing for Risk Adjustment** (*not required, but would provide additional support of adequacy of risk model, e.g., testing of risk model in another data set; sensitivity analysis for missing data; other methods that were assessed*)

N/A

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

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

**2b4.1. Describe the method for determining if statistically significant and clinically/practically meaningful differences in performance measure scores among the measured entities can be identified** (*describe the steps―do not just name a method; what statistical analysis was used? Do not just repeat the information provided related to performance gap in 1b)*

The method for discriminating hospital performance has not been determined. For public reporting of measures of hospital outcomes developed with similar methodology, CMS characterizes the uncertainty associated with the RSMR by estimating the 95% interval estimate. This is similar to a 95% confidence interval but is calculated differently. If the RSMR’s interval estimate does not include the national observed mortality rate (is lower or higher than the rate), then CMS is confident that the hospital’s RSMR is different from the national rate, and describes the hospital on the Hospital Compare website as “better than the U.S. national rate” or “worse than the U.S. national rate.” If the interval includes the national rate, then CMS describes the hospital’s RSMR as “no different than the U.S. national rate” or “the difference is uncertain.” CMS does not classify performance for hospitals that have fewer than 25 cases in the three-year period.

However, the decision to publicly report this hybrid AMI mortality measure and the approach to discriminating performance has not been determined.

During measure development, we assessed variation in AMI RSMRs among hospitals in the development dataset (**Dataset 1**) by examining the distribution of the hospital RSMRs.

**2b4.2. What were the statistical results from testing the ability to identify statistically significant and/or clinically/practically meaningful differences in performance measure scores across measured entities?** (e.g., *number and percentage of entities with scores that were statistically significantly different from mean or some benchmark, different from expected; how was meaningful difference defined*)

Analyses show substantial variation in RSMRs among hospitals. Using data from **Dataset 1** (development sample), the mean hospital RSMR was 10.8% with a range of 9.6% to 13.1%. The interquartile range was 10.3% - 11.1%. Using data from **Dataset 1** (validation sample), the mean hospital RSMR was 11.0% with a range of 7.7% to 15.8%. The interquartile range was 10.2% - 11.7%.

Note that this range is slightly narrower than what would be expected for a full national sample due to the self-selection of hospitals participating in Dataset 1 (development sample).

**2b4.3. What is your interpretation of the results in terms of demonstrating the ability to identify statistically significant and/or clinically/practically meaningful differences in performance across measured entities?** (i*.e., what do the results mean in terms of statistical and meaningful differences?*)

The variation in rates suggests there are meaningful differences across hospitals in the 30-day risk-standardized hybrid AMI mortality measure.

**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*)

We explicitly included only data elements that fulfilled the criteria for data element feasibility in this hybrid measure. Specifically, these criteria required that variables be:

1. Consistently obtained in the target population based on current clinical practice;

2. Captured with a standard definition and recorded in a standard format; and

3. Entered in structured fields that are feasibly retrieved from current EHR systems.

For the EHR data elements used in the measure’s risk models, we anticipate that there will be some missing data. However, we included only those variables that met these criteria and, therefore, anticipated that the overall rate of missing data elements would be low.

We examined rates of data capture and missing data in **Dataset 1** (development sample), as well as in the EHR data element feasibility and validity testing datasets (**Dataset 2, 3 and 4**).

During original development, the only data element that was missing at a meaningful rate was the hospital upper limit of normal for troponin. However, since completing validity testing using Dataset 3, and 4, hospitals have confirmed the ability to electronically capture and submit the upper limit of normal for troponin.

As was shown in Section 2b1.3, missing values were rare in this cohort. Because missing values were rare in the development and testing datasets, it was not necessary to do tests of bias in measure results. For those risk-adjustment variables that were missing, we imputed the median value of the sample for the continuous variables. No categorical variables were included in the final model. Due to the small amount of missing data, we do not expect that the missing data affected the measure score results.

All other data elements were found to be consistently and feasibly extracted from current EHRs. This is encouraging and indicates that missing data would have minimal effect on the measure calculation.

**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*)

We report the capture rate of all EHR data elements in each dataset in Section 2b1.3.

**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*)

The rate of missing values was low in all of the datasets and for all hospitals used for testing and therefore not likely to introduce bias. However, we did account for potential outlier values as well as missing values in our risk models to reduce any small possibility of bias. Approaches to handling missing clinical data in measure calculation will be reassessed during implementation.