

## NATIONAL QUALITY FORUM—Measure Testing (subcriteria 2a2, 2b2-2b7)

Last Updated 12/3/13

**Measure Number** (if previously endorsed): Click here to enter NQF number

**Measure Title:** Hospital 30-Day Risk-Standardized Acute Myocardial Infarction (AMI) Mortality  
eMeasure

**Date of Submission:** [12/23/2013](#)

**Type of Measure:**

<input type="checkbox"/> Composite – <b>STOP – use composite testing form</b>	<input checked="" type="checkbox"/> Outcome (including PRO-PM)
<input type="checkbox"/> Cost/resource	<input type="checkbox"/> Process
<input type="checkbox"/> Efficiency	<input type="checkbox"/> Structure

### Instructions

- Measures must be tested for all the data sources and levels of analyses that are specified. **If there is more than one set of data specifications or more than one level of analysis, contact NQF staff** about how to present all the testing information in one form.
- For **all** measures, sections 1, 2a2, 2b2, 2b3, and 2b5 must be completed.
- For **outcome and resource use** measures, section 2b4 also must be completed.
- If specified for **multiple data sources/sets of specifications** (e.g., claims and EHRs), section 2b6 also must be completed.
- Respond to **all** questions as instructed with answers immediately following the question. All information on testing to demonstrate meeting the subcriteria for reliability (2a2) and validity (2b2-2b6) must be in this form. An appendix for *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 20 pages (*including 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](#).

**Note:** The information provided in this form is intended to aid the Steering Committee and other stakeholders in understanding to what degree the testing results for this measure meet NQF's evaluation criteria for testing.

**2a2. Reliability testing** <sup>10</sup> 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 **PRO-PMs and composite performance measures**, reliability should be demonstrated for the computed performance score.

**2b2. Validity testing** <sup>11</sup> 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 **PRO-PMs and composite performance measures**, validity should be demonstrated for the computed performance score.

**2b3.** Exclusions are supported by the clinical evidence; otherwise, they are supported by evidence of sufficient

frequency of occurrence so that results are distorted without the exclusion; <sup>12</sup>

**AND**

If patient preference (e.g., informed decisionmaking) is a basis for exclusion, there must be evidence that the exclusion impacts performance on the measure; in such cases, the measure must be specified so that the information about patient preference and the effect on the measure is transparent (e.g., numerator category computed separately, denominator exclusion category computed separately). <sup>13</sup>

**2b4. For outcome measures and other measures when indicated** (e.g., resource use):

- **an evidence-based risk-adjustment strategy** (e.g., risk models, risk stratification) is specified; is based on patient factors that influence the measured outcome (but not factors related to disparities in care or the quality of care) and are present at start of care; <sup>14,15</sup> and has demonstrated adequate discrimination and calibration

**OR**

- rationale/data support no risk adjustment/ stratification.

**2b5. Data analysis of computed measure scores demonstrates that methods for scoring and analysis of the specified measure allow for identification of statistically significant and practically/clinically meaningful <sup>16</sup> differences in performance;**

**OR**

there is evidence of overall less-than-optimal performance.

**2b6. If multiple data sources/methods are specified, there is demonstration they produce comparable results.**

**2b7. For eMeasures, composites, and PRO-PMs** (or other measures susceptible to missing data), 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.

**12.** Examples of evidence that an exclusion distorts measure results include, but are not limited to: frequency of occurrence, variability of exclusions across providers, and sensitivity analyses with and without the exclusion.

**13.** Patient preference is not a clinical exception to eligibility and can be influenced by provider interventions.

**14.** Risk factors that influence outcomes should not be specified as exclusions.

- 15.** Risk models should not obscure disparities in care for populations by including factors that are associated with differences/inequalities in care, such as race, socioeconomic status, or gender (e.g., poorer treatment outcomes of African American men with prostate cancer or inequalities in treatment for CVD risk factors between men and women). It is preferable to stratify measures by race and socioeconomic status rather than to adjust out the differences.
- 16.** With large enough sample sizes, small differences that are statistically significant may or may not be practically or clinically meaningful. The substantive question may be, for example, whether a statistically significant difference of one percentage point in the percentage of patients who received smoking cessation counseling (e.g., 74 percent v. 75 percent) is clinically meaningful; or whether a statistically significant difference of \$25 in cost for an episode of care (e.g., \$5,000 v. \$5,025) is practically meaningful. Measures with overall less-than-optimal performance may not demonstrate much variability across providers.

## 1. DATA/SAMPLE USED FOR ALL TESTING OF THIS MEASURE

*Often the same data are used for all aspects of measure testing. In an effort to eliminate duplication, the first five questions apply to all measure testing. If there are differences by aspect of testing, (e.g., reliability vs. validity) be sure to indicate the specific differences in question 1.7.*

**1.1. What type of data was used for testing?** (Check all the sources of data identified in the measure specifications and data used for testing the measure. Testing must be provided for all the sources of data specified and intended for measure implementation. **If different data sources are used for the numerator and denominator, indicate N [numerator] or D [denominator] after the checkbox.**)

Measure Specified to Use Data From: (must be consistent with data sources entered in S.23)	Measure Tested with Data From:
<input type="checkbox"/> abstracted from paper record	<input type="checkbox"/> abstracted from paper record
<input type="checkbox"/> administrative claims	<input checked="" type="checkbox"/> administrative claims
<input type="checkbox"/> clinical database/registry	<input checked="" type="checkbox"/> clinical database/registry
<input type="checkbox"/> abstracted from electronic health record	<input checked="" type="checkbox"/> abstracted from electronic health record
<input checked="" type="checkbox"/> eMeasure (HQMF) implemented in EHRs	<input checked="" type="checkbox"/> eMeasure (HQMF) implemented in EHRs
<input type="checkbox"/> other: <a href="#">Click here to describe</a>	<input type="checkbox"/> other: <a href="#">Click here to describe</a>

**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).

ACTION Registry(R)-GWTG(TM) (AR-G), Medicare Part A claims, hospital EHR data, survey data

The dataset used varies by testing type; see Section 1.7 for details. Measure development used AR-G data merged with Medicare Part A claims data (AR-G-CMS). Testing of the eSpecified eMeasure used hospital electronic health record (EHR) data.

**1.3. What are the dates of the data used in testing?** 1/1/2009-10/1/2012

The dates used vary by testing type; see Section 1.7 for details.

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

Measure Specified to Measure Performance of: (must be consistent with levels entered in item S.26)	Measure Tested at Level of:
<input type="checkbox"/> individual clinician	<input type="checkbox"/> individual clinician
<input type="checkbox"/> group/practice	<input type="checkbox"/> group/practice
<input checked="" type="checkbox"/> hospital/facility/agency	<input checked="" type="checkbox"/> hospital/facility/agency
<input type="checkbox"/> health plan	<input type="checkbox"/> health plan
<input type="checkbox"/> other: Click here to describe	<input type="checkbox"/> other: Click here to describe

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

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

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

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

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

Please note this model was developed in merged registry and claims data and then eSpecified to create an eMeasure. We performed testing on both the registry model and the fully eSpecified eMeasure. We tested the eMeasure using EHR data.

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

Measure development dataset

The measure development dataset used merged AR-G registry and Medicare Part A claims data from 1/1/2009-12/31/2009. It included 280 hospitals and 20,540 AMI admissions for patients aged 65 years

Version 6.5 08/20/13

and older. The mortality rate in this dataset was 10.80%. The merged 2009 AR-G-CMS dataset was used for:

- Data element reliability testing (Section 2a2)
- Measure score validity testing (Section 2b2)
- Testing of measure risk adjustment (Section 2b4)
- Testing to identify meaningful differences in performance (Section 2b5)

#### Measure validation dataset

The measure validation dataset used merged AR-G registry and Medicare Part A claims data from 1/1/2010-12/31/2010. It included 460 hospitals and 34,196 AMI admissions for patients aged 65 and older. The mortality rate in this dataset was 10.98%. The merged 2010 AR-G-CMS dataset was used for:

- Data element reliability testing (Section 2a2)
- Testing of measure risk adjustment (Section 2b4)

#### EHR data

The EHR dataset included EHR data collected from 1/1/2011-10/1/2012. It included data collected from three hospitals using three different EHR vendors, and a total of 140 patients. Patients included in the sample were aged 65 years and older and were admitted as inpatients with a principal discharge diagnosis of AMI. The EHR dataset was used for:

- eMeasure data element validity testing (Section 2b2)

#### Measure exclusions testing dataset

The measure exclusions testing dataset used merged AR-G registry and Medicare Part A claims data from the 1/1/2009-12/31/2009 measure development dataset, before exclusions were applied. This dataset included 280 hospitals and 21,640 AMI admissions for patients aged 65 and older. The measure exclusions testing dataset was used for:

- Exclusions testing (Section 2b3)

#### Survey data

Survey data were collected from health information technology (IT) and quality experts from seven hospitals between 6/1/2012 and 8/1/2012. They were also collected from nine EHR vendors representing 85% of the current EHR market between 6/1/2012 and 8/1/2012. Survey data were used for:

- eMeasure usability testing
- eMeasure feasibility testing

---

## **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)

**Critical data elements used in the measure**

To assess the reliability of the risk-adjustment variables, we examined the temporal variation of the odds ratios and 95% confidence intervals of the model variables in the 2009 model development dataset vs. the 2010 model validation dataset.

**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)

**Critical data elements used in the measure**

**Table 1. Final logistic regression model odds ratios by dataset**

Description	2009 Development Sample	2010 Validation Sample
Age (years)	1.07 (1.06, 1.07)	1.07 (1.06, 1.07)
Heart Rate: HR<70 (10 bpm)	0.95 (0.88, 1.03)	0.99 (0.93, 1.05)
Heart Rate: HR≥70 (10 bpm)	1.16 (1.13, 1.19)	1.14 (1.12, 1.17)
Systolic Blood Pressure (10 mm Hg)	0.78 (0.76, 0.77)	0.78 (0.76, 0.79)
Troponin Ratio (ng/mL) (per 10 units)	1.13 (1.10, 1.15)	1.12 (1.10, 1.14)
Creatinine (mg/dL)	1.96 (1.82, 2.10)	1.85 (1.75, 1.95)

**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?)

**Critical data elements used in the measure**

The stability of the odds ratios from data element reliability testing (performed with registry data) indicates data elements are reliable.

---

**2b2. VALIDITY TESTING**

**2b2.1. What level of validity testing was conducted?** (may be one or both levels)

☒ **Critical data elements** (data element validity must address ALL critical data elements)

☒ **Performance measure score**

☒ **Empirical validity testing**

☐ **Systematic assessment of face validity of performance measure score as an indicator of quality**

or resource use (i.e., is an accurate reflection of performance on quality or resource use and can distinguish good from poor performance)

**2b2.2. For each level 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)

**Critical data elements**

Data element validity testing was done on the fully eSpecified eMeasure by comparing electronically extracted data elements to those manually abstracted at three hospitals, each with a different EHR vendor. Validity was assessed by determining the percent agreement between the electronically extracted and manually abstracted data element values. Inter-rater reliability (IRR) of the abstractors was also assessed to ensure that the manually abstracted data elements were an accurate standard against which the electronically extracted data elements could be compared.

**Performance measure score: Empirical validity testing**

To assess the validity of the measure score, we applied the model in the publicly reported claims-based AMI mortality measure in the study sample and calculated hospital risk-standardized mortality rates (RSMRs). Then we calculated the weighted Pearson correlation between the hospital RSMR based on the claims-based model and the hospital RSMR based on our final model.

It is important to note that the publicly reported claims-based AMI mortality measure was also previously validated with a comprehensive medical record model from an earlier time period. Specifically, claims-based model validation was conducted by building comparable models using abstracted medical record data for risk adjustment using Cooperative Cardiovascular Project data.

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

**Critical data elements**

We assessed data element validity of the fully eSpecified eMeasure using the percent agreement between findings of electronic extraction and manual abstraction in the EHR systems at three hospitals as follows:

**Table 2. Percentage of patients in the electronically extracted data that were eligible based on inclusion criteria, as identified by nurse abstraction**

	Hospital A (n=60)	Hospital B (n=40)	Hospital C (n=40)
Patients with a principal discharge diagnosis of AMI (%)	98.3	95	95
Patients with an inpatient admission (%)	100	100	92.5

<b>Table 3. Identification of transfer patients by electronic extraction and nurse abstraction</b>			
	<b>Hospital A (n=60)</b>	<b>Hospital B (n=40)</b>	<b>Hospital C (n=40)</b>
Identified as transfers in by both electronic extraction and nurse abstraction (%)	1.7	25.0	0.0
Identified as transfers in by electronic extraction only (%)	23.3	25.0	0.0
Identified as transfers in by nurse abstraction only (%)	10.0	5.0	2.5
Not identified as transfers in by either electronic or nurse abstraction (%)	65.0	45.0	97.5

<b>Table 4. Agreement between electronically extracted and manually abstracted risk-adjustment variables</b>			
	<b>Hospital A (n=53)</b>	<b>Hospital B (n=26)</b>	<b>Hospital C (n=34)</b>
Age	100.0	100.0	100.0
Heart Rate	98.1	100.0	91.2
Systolic Blood Pressure	98.1	100.0	88.2
Patient Troponin	98.1	100.0	94.1
Hospital Troponin	0.0*	100.0	0.0**
Creatinine	100.0	96.2	82.4
* = Electronic extraction did not produce this number      **= Electronic extraction produced incorrect values			

IRR values ranged from 92.9% to 99.1% across hospitals included in this testing, indicating a high level of agreement between the nurse abstractors' reviews at all hospitals.



**Performance measure score: Empirical validity testing**

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 eMeasure RSMRs and RSMRs based on the previously developed, publicly reported claims-based AMI mortality measure (hospital volume-weighted Pearson correlation coefficient=0.86)**



**2b2.4. What is your interpretation of the results in terms of demonstrating validity? (i.e., what do the results mean and what are the norms for the test conducted?)**

**Critical data elements**

Data element validity testing of the fully eSpecified AMI mortality eMeasure supported the overall validity of nearly all of the data elements included in the eMeasure. There were notable issues with identifying patients who had been transferred into the hospital. There were significant issues with extracting the hospital upper limit of normal for troponin, which we asked hospitals to provide manually. All other data elements for cohort identification and risk adjustment were consistently found for all patients and were both extractable and accurate.

**Performance measure score: Empirical validity testing**

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

## 2b3. EXCLUSIONS ANALYSIS

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

**2b3.1. Describe the method of testing exclusions and what it tests** (*describe the steps—do not just name a method; what was tested, e.g., whether exclusions affect overall performance scores; what statistical analysis was used*)

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

**2b3.2. What were the statistical results from testing exclusions?** (*include overall number and percentage of individuals excluded, frequency distribution of exclusions across measured entities, and impact on performance measure scores*)

Results are presented for the merged 2009 AR-G-CMS development dataset. For the purposes of tabulation, exclusions were performed sequentially. Thus, a hospital stay that would be excluded based on multiple criteria was counted in the first criterion only.

- 1) Discharged against medical advice (AMA) (n=53; 0.24%)
- 2) Transferred in (n=615; 2.8%)
- 3) Unknown death (records with missing vital status) in Medicare Enrollment Database (n=0; 0%)
- 4) Unreliable data (n=1; 0%)
- 5) Multiple AMI admissions in 2009 (n=431; 2.0%)

**2b3.3. What is your interpretation of the results in terms of demonstrating that exclusions are needed to prevent unfair distortion of performance results?** (*i.e., the value outweighs the burden of increased data collection and analysis. Note: If patient preference is an exclusion, the measure must be specified so that the effect on the performance score is transparent, e.g., scores with and without exclusion*)

The decision to exclude patients discharged AMA is based on clinical judgment to make the measure fair and is unlikely to distort the results given the very low frequency. Excluding patients transferring into a hospital does not actually exclude acute episodes from the measure, but considers the hospital that initially admits the patient as the one accountable for the outcome, avoiding double counting and clarifying accountability. The third and fourth exclusions are necessary for valid calculation of the measure; they affect very few patients. The final exclusion is to ensure that episodes are independent for statistical purposes.

## 2b4. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES

**If not an intermediate or health outcome, or PRO-PM, or resource use measure, skip to section [2b5](#).**

**2b4.1. 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](#)

**2b4.2. If an outcome or resource use measure is not risk adjusted or stratified, provide rationale and analyses to demonstrate that controlling for differences in patient characteristics (case mix) is not needed to achieve fair comparisons across measured entities.**

N/A

**2b4.3. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors used in the statistical risk model or for stratification by risk (e.g., potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of  $p < 0.10$ ; correlation of  $x$  or higher; patient factors should be present at the start of care and not related to disparities)**

The goal of risk adjustment is to account for different patient demographic and clinical characteristics at the time of admission (hospital case mix), enabling interpretation of any identified differences in quality. Candidate variables for risk adjustment were clinically relevant variables available in the AR-G dataset that were available at presentation and deemed feasible for use in an eMeasure. Conditions that may represent adverse outcomes due to care received during the index hospital stay are not included in the risk-adjustment model. We assessed a variable's eMeasure feasibility according to the following three criteria.

To be included in the model, the data element must 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.
3. Entered in structured fields that are feasibly retrieved from current EHR systems.

Only variables that met all three criteria were considered for inclusion. We used a modified approach to stepwise regression that takes both clinical and statistical considerations into account in selecting final variables.

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

To create a model with increased usability while retaining excellent model performance, we tested the performance of the model without those variables considered to be questionably feasible. 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 eMeasure-feasible.

During model development using the merged 2009 AR-G-CMS data, 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. We summarized the model information of all 1,000 iterations on the following: number and frequency of times that a variable is selected (e.g., 70% would mean that the candidate variable was selected as significant at  $p < 0.05$  in 70% of the times), minimum, maximum, and the range of the standardized coefficient for a selected variable. We also assessed the direction and magnitude of the distribution of regression coefficients.

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 eMeasure feasibility (see 2b4.3) and were excluded from the final model. The following variables were selected for inclusion in the final model:

- Age (years)
- Heart rate (bpm)
- Systolic blood pressure (mmHg)
- Troponin ratio (initial troponin value (ng/mL) / initial troponin URL (ng/mL))
- Creatinine (mg/dL)

**2b4.5. Describe the method of testing/analysis used to develop and validate the adequacy of the statistical model or stratification approach** (*describe the steps—do not just name a method; what statistical analysis was used*)

We used a merged 2009 AR-G-CMS dataset to develop the risk-adjusted model. We validated the model using a merged 2010 AR-G-CMS dataset. Model performance was assessed in both datasets. Due to the low sample size, we did not split the development sample in our assessment of the risk-adjusted model. The approach to assessing model performance is as follows:

For both datasets, we computed the following summary statistics for assessing model performance (Harrell and Shih, 2001):

- (1) Over-fitting indices (over-fitting refers to the phenomenon in which a model accurately describes the relationship between predictive variables and outcome in the dataset used for development but fails to provide valid predictions in new patients)
- (2) Predictive ability
- (3) Area under the receiver operating characteristic (ROC) curve
- (4) Distribution of residuals

We also examined the odds ratios and frequencies of the final model variables in two separate years of data.

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.

*Provide the statistical results from testing the approach to controlling for differences in patient characteristics (case mix) below.*

**If stratified, skip to [2b4.9](#)**

**2b4.6. 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.

Table 5. Model Performance: Discrimination Results Based on the Logistic Regression Model		
Indices	2009 Development Sample	2010 Validation Sample
Number of Admissions	20,540	34,196
C-Statistic	0.78	0.78
Predictive Ability (lowest decile %, highest decile %)	1.2, 37.5	1.2, 37.4

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

Table 6. Model Performance: Calibration Results Based on the Logistic Regression Model		
Indices	2009 Development Sample	2010 Validation Sample
Number of Admissions	20,540	34,196
Calibration		
$\gamma_0, \gamma_1$	–	-0.013, 0.979
Residuals Lack of Fit (Pearson Residual Fall %)		
<-2	0.015	0.000
[-2, 0)	89.187	89.019
[0, 2)	4.869	4.849
[2+	5.930	6.132

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

Table 7. 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

#### 2b4.9. Results of Risk Stratification Analysis:

N/A

**2b4.10. What is your interpretation of the results in terms of demonstrating adequacy of controlling for differences in patient characteristics (case mix)?** (i.e., *what do the results mean and what are the norms for the test conducted*)

The C-statistic of 0.78 indicates excellent model discrimination. The calibration value of close to 0 at one end and close to 1 to the other end indicates good calibration of the model. The risk decile plot shows excellent discrimination of the model and good predictive ability.

**2b4.11. Optional Additional Testing for Risk Adjustment** (*not required, but would provide additional support of adequacy of risk model, e.g., testing of risk model in another data set; sensitivity analysis for missing data; other methods that were assessed*)

N/A

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

**2b5.1. Describe the method for determining if statistically significant and clinically/practically meaningful differences in performance measure scores among the measured entities can be identified** (*describe the steps—do not just name a method; what statistical analysis was used? Do not just repeat the information provided related to performance gap in 1b*)

For nine currently publicly reported measures of hospital outcomes, CMS estimates an interval estimate for each risk-standardized rate to characterize the amount of uncertainty associated with the rate. It then compares the interval estimate to the national crude rate for the outcome and categorizes hospitals as “better than,” “worse than,” or “no different than” the U.S. national rate. However, the decision to publicly report this AMI mortality eMeasure and the approach to discriminating performance has not been determined.

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

**2b5.2. What were the statistical results from testing the ability to identify statistically significant and/or clinically/practically meaningful differences in performance measure scores across measured entities?** (e.g., number and percentage of entities with scores that were statistically significantly different from mean or some benchmark, different from expected; how was meaningful difference defined)

In 2009, the mean hospital RSMR was 10.8%, with a range of 9.6% to 13.1%. The interquartile range was 10.3% to 11.1%. In 2010, the mean hospital RSMR was 11.0%, with a range of 7.7% to 15.8%. The interquartile range was 10.2% to 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 the AR-G.

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

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

---

**2b6. COMPARABILITY OF PERFORMANCE SCORES WHEN MORE THAN ONE SET OF SPECIFICATIONS**  
***If only one set of specifications, this section can be skipped.***

**Note:** This criterion is directed to measures with more than one set of specifications/instructions (e.g., one set of specifications for how to identify and compute the measure from medical record abstraction and a different set of specifications for claims or eMeasures). It does not apply to measures that use more than one source of data in one set of specifications/instructions (e.g., claims data to identify the denominator and medical record abstraction for the numerator). ***If comparability is not demonstrated, the different specifications should be submitted as separate measures.***

**2b6.1. Describe the method of testing conducted to demonstrate comparability of performance scores for the same entities across the different data sources/specifications** (describe the steps—do not just name a method; what statistical analysis was used)

**2b6.2. What were the statistical results from testing comparability of performance scores for the same entities when using different data sources/specifications?** (e.g., correlation, rank order)

**2b6.3. What is your interpretation of the results in terms of demonstrating comparability of performance measure scores for the same entities across the different data sources/specifications?**

(i.e., what do the results mean and what are the norms for the test conducted)

---

## **2b7. MISSING DATA ANALYSIS AND MINIMIZING BIAS**

**2b7.1. 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*)

### **Measure development dataset**

Missing values were rare in this cohort. There were no missing data for age, 0.14% missing for heart rate, 0.15% for systolic blood pressure, 1.71% for troponin ratio, and 0.49% for creatinine based on the 2009 NCDR AR-G dataset. 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.

### **EHR data**

We explicitly included only data elements that fulfilled our criteria for data element feasibility in the eMeasure. 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.
3. Entered in structured fields that are feasibly retrieved from current EHR systems.

Because we included only those variables that met these criteria, the overall rate of missing data elements was low.

The only data element that was missing at a meaningful rate for the three hospitals during data element validity testing (see Section 2b2) was the hospital upper limit of normal for troponin. However, we do not expect this to affect the measure calculation because this value can be provided manually by hospitals.

All other data elements were found to be consistently and feasibly extracted from current EHRs. More comprehensive testing for missing data could not be conducted using the eSpecified eMeasure due to the lack of a nationally representative EHR dataset at the time of measure development. However, the minimal amount of missing data at the three test hospitals is encouraging and indicates that missing data would have minimal effect on the measure calculation.



We expect that during implementation of the hybrid eMeasure, we would impute the median value of the sample for any missing continuous risk-adjustment variables. Again, due to the small observed amount of missing data, we do not expect that this approach will affect or bias the measure results.

#### **Measure development dataset**

Missing values were rare in this cohort. There were no missing data for age, 0.14% missing for heart rate, 0.15% for systolic blood pressure, 1.71% for troponin ratio, and 0.49% for creatinine based on the 2009 NCDR AR-G dataset.

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.

#### **EHR data**

We explicitly included only data elements that fulfilled our criteria for data element feasibility in the eMeasure. 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.
3. Entered in structured fields that are feasibly retrieved from current EHR systems.

Because we included only those variables that met these criteria, the overall rate of missing data elements was low.

The only data element that was missing at a meaningful rate for the three hospitals during data element validity testing (see Section 2b2) was the hospital upper limit of normal for troponin. However, we do not expect this to affect the measure calculation because this value can be provided manually by hospitals.

All other data elements were found to be consistently and feasibly extracted from current EHRs. More comprehensive testing for missing data could not be conducted using the eSpecified eMeasure due to the lack of a nationally representative EHR dataset at the time of measure development. However, the minimal amount of missing data at the three test hospitals is encouraging and indicates that missing data would have minimal effect on the measure calculation.

We expect that during implementation of the hybrid eMeasure, we would impute the median value of the sample for any missing continuous risk-adjustment variables. Again, due to the small observed amount of missing data, we do not expect that this approach will affect or bias the measure results.

**2b7.2. 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)

**Measure development dataset**

More detailed testing was not conducted because the amount of missing data element values was low (<2%), as detailed in question 2b7.1.

**EHR data**

Table 8 shows the percent of data element values that were missing in the three hospitals where field testing was conducted.

**Table 8. Percent of variables missing from the EHR data**

	Hospital A (n=53)	Hospital B (n=26)	Hospital C (n=34)
Age	0	0	0
AMI Status	0	0	0
Transferred In	0	0	0
Heart Rate	1.89	0	0
Systolic Blood Pressure	1.89	0	0
Patient Troponin	0	0	0
Hospital Troponin	100	0	0
Creatinine	0	0	0

As mentioned previously, the only data element that was found to be missing often among the three hospitals was the hospital upper limit of normal for troponin. There were significant issues with extracting this data element from the EHR data. However, since hospitals can provide this value manually, we do not expect this to affect the measure calculation.

**2b7.3. 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*)

**Measure development dataset**

N/A. More detailed testing was not conducted because the amount of missing data element values was low (<2%), as detailed in question 2b7.1.

**EHR data**

Based on the results of our testing, we expect the amount of missing data from the EHR to be low. We plan to impute the median value of the sample for missing risk-adjustment variables during implementation of the hybrid eMeasure. For missing values for the hospital upper limit of normal for troponin, we plan to use the value that hospitals provide manually. Due to the small observed amount of missing data, we do not expect that this approach will affect or bias the measure results.