NATIONAL QUALITY FORUM

MEASURE WORKSHEET

This document summarizes the evaluation of the measure as it progresses through NQF's Consensus Development Process (CDP). The information submitted by measure developers/stewards is included after the Brief Measure Information, Preliminary Analysis, and Pre-meeting Public and Member Comments sections.

To navigate the links in the worksheet: Click to go to the link. ALT + LEFT ARROW to return

Purple text represents the responses from measure developers.

Red text denotes developer information that has changed since the last measure evaluation review.

Brief Measure Information

NQF #: 0535

Measure Title: 30-day all-cause risk-standardized mortality rate following percutaneous coronary intervention (PCI) for patients without ST segment elevation myocardial infarction (STEMI) and without cardiogenic shock

Measure Steward: American College of Cardiology

Brief Description of Measure: This measure estimates hospital risk-standardized 30-day all-cause mortality rate following percutaneous coronary intervention (PCI) among patients who are 18 years of age or older without STEMI and without cardiogenic shock at the time of procedure. The measure uses clinical data available in the National Cardiovascular Data Registry (NCDR) CathPCI Registry for risk adjustment. For the purpose of development and testing, the measure used a Medicare fee-for-service (FFS) population of patients 65 years of age or older with a PCI. For the purpose of maintenance, we tested the performance of the measure in a cohort of patients whose vital status was determined from the National Death Index. As such it reflects an all-payor sample as opposed to only the Medicare population. This is consistent with the measure's intent to be applicable to the full population of PCI patients.

Developer Rationale: This measure will describe hospital-level mortality rates following PCI in patients without STEMI and without cardiogenic shock, with the overriding goal to reduce 30-day mortality rates to best-in-class. The expectation is that providing this information to hospitals, coupled with public reporting of hospitals' results, will drive internal hospital quality improvement efforts to focus efforts on reducing PCI mortality. Of note, the measure includes not only in-hospital deaths, but also deaths occurring after hospital discharge. This perspective may motivate hospitals to look for opportunities not only within the organization, but to better coordinate the transition of care from the inpatient to the outpatient arena.

Numerator Statement: The outcome for this measure is all–cause death within 30 days following a PCI procedure in patients without STEMI and without cardiogenic shock at the time of the procedure.

Denominator Statement: The target population for this measure includes inpatient and outpatient hospital stays with a PCI procedure for patients at least 18 years of age, without STEMI and without cardiogenic shock at the time of procedure.

Denominator Exclusions: Hospital stays are excluded from the cohort if they meet any of the following criteria:

(1) PCIs that follow a prior PCI in the same admission (either at the same hospital or a PCI performed at another hospital prior to transfer).

This exclusion is applied in order to avoid assigning the death to two separate admissions.

(2) For patients with inconsistent or unknown vital status or other unreliable data (e.g. date of death precedes date of PCI);

(3) Subsequent PCIs within 30-days. The 30-day outcome period for patients with more than one PCI may overlap. In order to avoid attributing the same death to more than one PCI (i.e. double counting a single patient death), additional PCI procedures within 30 days of the death are not counted as new index procedures.

(4) PCIs for patients with more than 10 days between date of admission and date of PCI. Patients who have a PCI after having been in the hospital for a prolonged period of time are rare and represent a distinct population that likely has risk factors related to the hospitalization that are not well quantified in the registry.

Measure Type: Outcome

Data Source: Claims, Other, Registry Data

Level of Analysis: Facility, Other

IF Endorsement Maintenance – Original Endorsement Date: Aug 05, 2009 **Most Recent Endorsement Date:** Sep 08, 2014

Staff Preliminary Analysis: Maintenance of Endorsement

To maintain NQF endorsement endorsed measures are evaluated periodically to ensure that the measures still meets the NQF endorsement criteria ("maintenance"). The emphasis for maintaining endorsement is focused on how effective the measure is for promoting improvements in quality. Endorsed measures should have some experience from the field to inform the evaluation. The emphasis for maintaining endorsement is noted for each criterion.

Criteria 1: Importance to Measure and Report

1a. Evidence

Maintenance measures – less emphasis on evidence unless there is new information or change in evidence since the prior evaluation.

1a. Evidence. The evidence requirements for a health outcome measure include providing empirical data that demonstrate a relationship between the outcome and at least one healthcare structure, process, intervention, or service; if these data not available, data demonstrating wide variation in performance, assuming the data are from a robust number of providers and results are not subject to systematic bias. For measures derived from patient report, evidence also should demonstrate that the target population values the measured outcome, process, or structure and finds it meaningful.

Summary of prior review in 2014

- In their rationale, the developer referenced literature supporting an association with improved survival and the use of preprocedural clopidogrel and glycoprotein 2b/3a inhibitors; the volume of iodinated contrast; and participation in continuous quality improvement programs.

Changes to evidence from last review

☒ **The developer attests that there have been no changes in the evidence since the measure was last evaluated.**

☐ **The developer provided updated evidence for this measure:**

Updates: The developer stated that there are no updates to the evidence.

The developer provided performance data from 1,365 hospitals and 1,127,423 admissions from 2011-2014 demonstrating a variation in risk-standardized mortality rates with a mean of 1.07% and a range from 0.51% to 2.70%.

Question for the Committee:

- *Does the stated rationale link lower mortality rates after PCI to at least one healthcare action?*

- *Is the performance data sufficient, in size and variance, to demonstrate that some hospitals are engaging in quality improvement activities to decrease mortality after PCI better than others?*

Guidance from the Evidence Algorithm

Health outcome measure (Box 1) -> relationship between the measured health outcome and at least one healthcare action is demonstrated -> Pass

Preliminary rating for evidence: ☒ **Pass** ☐ **No Pass**

1b. Gap in Care/Opportunity for Improvement and 1b. Disparities

Maintenance measures – increased emphasis on gap and variation

1b. Performance Gap. The performance gap requirements include demonstrating quality problems and opportunity for improvement.

- The developer provided data for the risk standardized mortality rates for all payers and all ages (>18 years) using National Cardiovascular Data Registry (NCDR) CathPCI data linked with National Death Index (NDI) from 2011-2014. From this linked data, 1,365 hospitals and 1,127,423 admissions were included. Two months of data were excluded due to missing data (October 2012 and November 2012).
- The developer provided the following RSMRs data:
    - Mean: 1.07%
    - Standard Deviation: 0.30%
    - Range: 0.51% to 2.70%
    - Interquartile Range: 0.87% to 1.24%
- The range of performance of the data are:

Percentile of RSMR	Mean RSMR
100% Max	0.0270
99%	0.0211
95%	0.0163
90%	0.0147
75% Q3	0.0124
50% Median	0.0104
25% Q1	0.0087
10%	0.0073
5%	0.0067
1%	0.0057
0% Min	0.0051

- The range of performance data by year are:

Percentile of RSMR	2011-2012	2012-2013	2013-2014
N	350,941	345,929	430,553
Mean	0.0100	0.0112	0.0113
Standard Deviation	0.0029	0.0034	0.0030
100% Max	0.0342	0.0332	0.0262
99%	0.0178	0.0221	0.0206
95%	0.0154	0.0175	0.0168
90%	0.0137	0.0154	0.0151
75% Q3	0.0115	0.0129	0.0129
50% Median	0.0095	0.0106	0.0108
25% Q1	0.0082	0.0089	0.0091
10%	0.0068	0.0076	0.0079
5%	0.0061	0.0069	0.0072
1%	0.0049	0.0060	0.0061
0% Min	0.0034	0.0045	0.0052

Disparities

- The developer provided the following hospital-level RSMR disparities data by race and hospital safety net status:

  Distribution of 30-day RSMR for No STEMI/No Shock Stratified by Quartile of Non-White Patients from 2011-2014

Description	RSMRs by Hospital Quartile of Non-White Patients			
	Q1	Q2	Q3	Q4
N	341	341	342	341
Mean	0.0112	0.0109	0.0112	0.0111
Standard Deviation	0.0027	0.0027	0.0028	0.0030
100% Max	0.0270	0.0227	0.0232	0.0245
50% Median	0.0107	0.0105	0.0107	0.0104
0% Min	0.0060	0.0051	0.0058	0.0056

	Safety Net Hospitals	Non-Safety Net Hospitals
Median RSMR	1.08%	1.05%
Interquartile Range	0.95% to 1.31%	0.93% to 1.23%

Questions for the Committee:

- *Does the measure demonstrate a quality problem related to mortality in patients undergoing PCI?*
- *Is a national performance measure still warranted?*

- *Are you aware of evidence that other disparities exist in this area of healthcare?*

Preliminary rating for opportunity for improvement: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee Pre-evaluation Comments: Criteria 1: Importance to Measure and Report (including 1a, 1b, 1c)

1a. Evidence:

- There is direct evidence that certain actions are associated with survival. Among these are the use of preprocedural clopidogrel and glycoprotein 2b/3a inhibitors; the volume of iodinated contrast; and participation in continuous quality improvement programs.
- The lack of new evidence since the last review in 2014 is concerning. There appear to be issues with timeliness of data collection for this measure.
- The developers indicate that the evidence is based upon "empirical data" rather than on randomized controlled clinical trials, but given the mortality as an endpoint then rcts would not be ethical, leaving us with empirical evidence as the highest level of evidence.
- This is an important measure, but the title of the measure if somewhat misleading. This measure only truly describes the mortality in the Medicare population. Therefore, I think the measure title should be changed as it may lead one to believe that this is the total mortality for the whole PCI population whereas, in fact, only the Medicare population is reported.
- Evidence exists to support that procedural quality, intra-procedural care, and post-procedural care can impact PCI outcomes.
- This is a maintenance measure and the developer states that these have been no changes to the evidence since the last NQF endorsement.

1b. Gap in Care/Opportunity for Improvement and 1b. Disparities:

- NO – current data were not provided. Data were only provided through 2014. Also ". Two months of data were excluded due to missing data (October 2012 and November 2012). Why would two months of data be missing?
- The developer provided the following RSMRs data:
    - Mean: 1.07%
    - Standard Deviation: 0.30%
    - Range: 0.51% to 2.70%
    - Interquartile Range: 0.87% to 1.24%
    - Safety net hospitals have marginally higher mortality rates
- Current performance data was not provided.
- The developer has evidence to show a gap in care between 2011 and 2014 that is rather consistent year over year.
- Racial disparities are not reported although they are known to vary considerably in this domain of care as indicated above, the mortality Medicare vs. non Medicare is not present, hence, the measure title is inaccurate
- 5X gap in 30-day mortality exists between best/worst performing centers.
- The interquartile range of the most recent data is narrow - 0.91-1.29 among a population of 430,553. While this is narrow, it represents a substantial number of deaths.

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability: Specifications **and** Testing
2b. Validity: Testing; Exclusions**;** Risk-Adjustment**;** Meaningful Differences**;** Comparability**;** Missing Data

Reliability

2a1. Specifications requires the measure, as specified, to produce consistent (reliable) and credible (valid) results about the quality of care when implemented. For maintenance measures – no change in emphasis – specifications should be evaluated the same as with new measures.

2a2. Reliability testing demonstrates if 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 enough to distinguish differences in performance across providers. For maintenance measures – less emphasis if no new testing data provided.

Validity

2b2. Validity testing should demonstrate the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality. For maintenance measures – less emphasis if no new testing data provided.

2b2-2b6. Potential threats to validity should be assessed/addressed.

Composite measures only:

2d. Empirical analysis to support composite construction. Empirical analysis should demonstrate that the component measures add value to the composite and that the aggregation and weighting rules are consistent with the quality construct.

Complex measure evaluated by Scientific Methods Panel? ☐ **Yes** ☒ **No**

Evaluators: NQF Staff

Evaluation of Reliability and Validity (and composite construction, if applicable): Link to Scientific Acceptability

Questions for the Committee regarding reliability:

- *Do you have any concerns that the measure can be consistently implemented (i.e., are measure specifications adequate)?*
- *Staff are satisfied with the reliability testing for the measure. Does the Committee think there is a need to discuss and/or vote on reliability?*

Questions for the Committee regarding validity:

- *Do you have any concerns regarding the validity of the measure (e.g., exclusions, risk-adjustment approach, etc.)?*
- *Staff are satisfied with the validity analyses for the measure. Does the Committee think there is a need to discuss and/or vote on validity?*

Preliminary rating for reliability:	☐ **High**	☒ **Moderate**	☐ **Low**	☐ **Insufficient**
Preliminary rating for validity:	☐ **High**	☒ **Moderate**	☐ **Low**	☐ **Insufficient**

Scientific Acceptability

Measure Number: 0535

Measure Title: 30-day all-cause risk-standardized mortality rate following percutaneous coronary intervention (PCI) for patients without ST segment elevation myocardial infarction (STEMI) and without cardiogenic shock

Staff Scientific Acceptability: Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. Measures must be judged to meet the subcriteria for both reliability and validity to pass this criterion

Instructions for filling out this form:

- Please complete this form for each measure you are evaluating.
- Please pay close attention to the skip logic directions. ***Directives that require you to skip questions are marked in red font.***
- If you are unable to check a box, please highlight or shade the box for your response.
- You must answer the "overall rating" item for both Reliability and Validity. Also, be sure to answer the composite measure question at the end of the form if your measure is a composite.
- For several questions, we have noted which sections of the submission documents you should ***REFERENCE*** and provided ***TIPS*** to help you answer them.
- ***It is critical that you explain your thinking/rationale if you check boxes that require an explanation.*** Please add your explanation directly below the checkbox in a different font color***.*** Also, feel free to add additional explanation, even if you select a checkbox where an explanation is not requested (if you do so, please type this text directly below the appropriate checkbox).
- ***Please refer to the*** Measure Evaluation Criteria and Guidance document ***(pages 18-24) and the 2-page*** Key Points document ***when evaluating your measures***. This evaluation form is an adaptation of Alogorithms 2 and 3, which provide guidance on rating the Reliability and Validity subcriteria.
- ***Remember*** that testing at either the data element level **OR** the measure score level is accepted for some types of measures, but not all (e.g., instrument-based measures, composite measures), and therefore, the embedded rating instructions may not be appropriate for all measures.
- ***Please base your evaluations solely on the submission materials provided by developers.*** NQF strongly discourages the use of outside articles or other resources, even if they are cited in the submission materials. If you require further information or clarification to conduct your evaluation, please communicate with NQF staff (methodspanel@qualityforum.org).

RELIABILITY

1. Are submitted specifications precise, unambiguous, and complete so that they can be consistently implemented?

   REFERENCE: "MIF_xxxx" document

   NOTE*: NQF staff will conduct a separate, more technical, check of eMeasure (eCQM) specifications, value sets, logic, and feasibility, so no need to consider these in your evaluation.*

   TIPS*: Consider the following: Are all the data elements clearly defined? Are all appropriate codes included? Is the logic or calculation algorithm clear? Is it likely this measure can be consistently implemented?*

   ☒Yes (go to Question #2)

   ☐No (please explain below, and go to Question #2) NOTE that even though ***non-precise specifications should result in an overall LOW rating for reliability***, we still want you to look at the testing results.

2. Was empirical reliability testing (at the data element or measure score level) conducted using statistical tests with the measure as specified?

   REFERENCE: "MIF_xxxx" document for specifications, testing attachment questions 1.1-1.4 and section 2a2

   TIPS*: Check the "NO" box below if: only descriptive statistics are provided; only describes process for data management/cleaning/computer programming; testing does not match measure specifications (i.e., data source, level of analysis, included patients, etc.)*

   ☒Yes (go to Question #3)

☐No, there is reliability testing information, but *not* using statistical tests and/or not for the measure as specified **OR** there is no reliability testing (please explain below, skip Questions #3-8, then go to Question #9)

3. Was reliability testing conducted with computed performance measure scores for each measured entity?

   REFERENCE: "Testing attachment_xxx", section 2a2.1 and 2a2.2

   ***TIPS**: Answer no if: only one overall score for all patients in sample used for testing patient-level data*

   ☒Yes (go to Question #4)

   ☐No (skip Questions #4-5 and go to Question #6)

4. Was the method described and appropriate for assessing the proportion of variability due to real differences among measured entities? *NOTE: If multiple methods used, at least one must be appropriate.*

   REFERENCE: Testing attachment, section 2a2.2

   ***TIPS**: Examples of appropriate methods include signal-to-noise analysis (e.g. Adams/RAND tutorial); random split-half correlation; other accepted method with description of how it assesses reliability of the performance score.*

   ☒Yes (go to Question #5)

   ☐No (please explain below, then go to question #5 and rate as INSUFFICIENT)

5. **RATING (score level)** - What is the level of certainty or confidence that the performance measure scores are reliable?

   REFERENCE: Testing attachment, section 2a2.2

   ***TIPS**: Consider the following: Is the test sample adequate to generalize for widespread implementation? Do the results demonstrate sufficient reliability so that differences in performance can be identified?*

   ☐High (go to Question #6)

   ☒Moderate (go to Question #6)

   ☐Low (please explain below then go to Question #6)

   ☐Insufficient (go to Question #6)

6. Was reliability testing conducted with patient-level data elements that are used to construct the performance measure?

   REFERENCE: Testing attachment, section 2a2.

   ***TIPS**: Prior reliability studies of the same data elements may be submitted; if comparing abstraction to "authoritative source/gold standard" go to Question #9)*

   ☒Yes (go to Question #7)

   ☐No (if there is score-level testing that you rated something other than INSUFFICIENT in Question #5, skip questions #7-9, then go to Question #10 (OVERALL RELIABILITY); otherwise, skip questions #7-8 and go to Question #9)

7. Was the method described and appropriate for assessing the reliability of ALL critical data elements?

   REFERENCE: Testing attachment, section 2a2.2

   ***TIPS**: For example: inter-abstractor agreement (ICC, Kappa); other accepted method with description of how it assesses reliability of the data elements*

   Answer no if: only assessed percent agreement; did not assess separately for all data elements (at least numerator, denominator, exclusions)

   ☒Yes (go to Question #8)

   ☐No (if no, please explain below, then go to Question #8 and rate as INSUFFICIENT)

8. **RATING (data element)** – Based on the reliability statistic and scope of testing (number and representativeness of patients and entities), what is the level of certainty or confidence that the data used in the measure are reliable?

   REFERENCE: Testing attachment, section 2a2

***TIPS**: Consider the following: Is the test sample adequate to generalize for widespread implementation? Can data elements be collected consistently?*

☒Moderate (skip Question #9 and go to Question #10 (OVERALL RELIABILITY); if score-level testing was NOT conducted, rate Question #10 as MODERATE)

☐Low (skip Question #9 and go to Question #10 (OVERALL RELIABILITY); if score-level testing was NOT conducted, rate Question #10 (OVERALL RELIABILITY) as LOW)

☐Insufficient (go to Question #9)

9. Was **empirical VALIDITY testing** of patient-level data conducted?

REFERENCE: testing attachment section 2b1.

NOTE: Skip this question if empirical reliability testing was conducted and you have rated Question #5 and/or #8 as anything other than INSUFFICIENT)

TIP:** You should answer this question ONLY if score-level or data element reliability testing was NOT conducted or if the methods used were NOT appropriate. For most measures, NQF will accept data element validity testing in lieu of reliability testing—but **check with NQF staff before proceeding, to verify.

☐Yes (go to Question #10 and answer using your rating from data element validity testing – Question #23)

☐No (please explain below, go to Question #10 (OVERALL RELIABILITY) and rate it as INSUFFICIENT. Then go to Question #11.)

OVERALL RELIABILITY RATING

10. **OVERALL RATING OF RELIABILITY** taking into account precision of specifications (see Question #1) and all testing results:

☐High (NOTE: Can be HIGH only if score-level testing has been conducted)

☒Moderate (NOTE: Moderate is the highest eligible rating if score-level testing has not been conducted)

☐Low (please explain below) [NOTE: Should rate LOW if you believe specifications are NOT precise, unambiguous, and complete]

☐Insufficient (please explain below) [NOTE: For most measure types, testing at both the score level and the data element level is not required, but check with NQF staff]

VALIDITY

ASSESSMENT OF THREATS TO VALIDITY

11. Were potential threats to validity that are relevant to the measure empirically assessed ()?

REFERENCE: Testing attachment, section 2b2-2b6

***TIPS**: Threats to validity that should be assessed include: exclusions; need for risk adjustment; ability to identify statistically significant and meaningful differences; multiple sets of specifications; missing data/nonresponse.*

☒Yes (go to Question #12)

☐No (please explain below and then go to Question #12) [NOTE that ***non-assessment of applicable threats should result in an overall INSUFFICENT rating for validity***]

12. Analysis of potential threats to validity: Any concerns with measure exclusions?

REFERENCE: Testing attachment, section 2b2.

***TIPS**: Consider the following: Are the exclusions consistent with the evidence? Are any patients or patient groups inappropriately excluded from the measure? Are the exclusions/exceptions of sufficient frequency and variation across providers to be needed (and outweigh the data collection burden)? If patient preference (e.g., informed*

decisionmaking) is a basis for exclusion, does it impact performance and if yes, is the measure specified so that the information about patient preference and the effect on the measure is transparent?

☐Yes (please explain below then go to Question #13)

☒No (go to Question #13)

☐Not applicable (i.e., there are no exclusions specified for the measure; go to Question #13)

13. Analysis of potential threats to validity: Risk-adjustment (this applies to all outcome, cost, and resource use measures and "NOT APPLICABLE" is not an option for those measures; the risk-adjustment questions (13a-13c, below) also may apply to other types of measures)

REFERENCE: Testing attachment, section 2b3.

13a. Is a conceptual rationale for social risk factors included? ☒Yes ☐No

13b. Are social risk factors included in risk model? ☐Yes ☒No

13c. Any concerns regarding the risk-adjustment approach?

***TIPS**: Consider the following: **If measure is risk adjusted**: If the developer asserts there is **no conceptual basis** for adjusting this measure for social risk factors, do you agree with the rationale? Are the candidate and final variables included in the risk adjustment model adequately described for the measure to be implemented? Are the candidate and final variables included in the risk adjustment model adequately described for the measure to be implemented? Are all of the risk adjustment variables present at the start of care (if not, do you agree with the rationale)? If social risk factors are not included in the risk-adjustment approach, do you agree with the developer's decision? Is an appropriate risk-adjustment strategy included in the measure (e.g., adequate model discrimination and calibration)? Are all statistical model specifications included, including a "clinical model only" if social risk factors are included in the final model? If a measure is NOT risk-adjusted, is a justification for **not risk adjusting** provided (conceptual and/or empirical)? Is there any evidence that contradicts the developer's rationale and analysis for not risk-adjusting?*

☐Yes (please explain below then go to Question #14)

☒No (go to Question #14)

☐Not applicable (e.g., this is a structure or process measure that is not risk-adjusted; go to Question #14)

14. Analysis of potential threats to validity: Any concerns regarding ability to identify meaningful differences in performance or overall poor performance?

REFERENCE: Testing attachment, section 2b4.

☒Yes (please explain below then go to Question #15)

☐No (go to Question #15)

The developer stated that the decision to publicly report this measure and approach to discriminating performance has not been determined, which may raise concerns on performance of the measure.

15. Analysis of potential threats to validity: Any concerns regarding comparability of results if multiple data sources or methods are specified?

REFERENCE: Testing attachment, section 2b5.

☐Yes (please explain below then go to Question #16)

☐No (go to Question #16)

☒Not applicable (go to Question #16)

16. Analysis of potential threats to validity: Any concerns regarding missing data?

REFERENCE: Testing attachment, section 2b6.

☐Yes (please explain below then go to Question #17)

☒No (go to Question #17)

ASSESSMENT OF MEASURE TESTING

17. Was empirical validity testing conducted using the measure as specified and with appropriate statistical tests?

    REFERENCE: Testing attachment, section 2b1.

    ***TIPS**: Answer no if: only face validity testing was performed; only refer to clinical evidence; only descriptive statistics; only describe process for data management/cleaning/computer programming; testing does not match measure specifications (i.e. data, eMeasure, level, setting, patients).*

    ☒Yes (go to Question #18)

    ☐No (please explain below, then skip Questions #18-23 and go to Question #24)

18. Was validity testing conducted with computed performance measure scores for each measured entity?

    REFERENCE: Testing attachment, section 2b1.

    ***TIPS**: Answer no if: one overall score for all patients in sample used for testing patient-level data.*

    ☒Yes (go to Question #19)

    ☐No (please explain below, then skip questions #19-20 and go to Question #21)

19. Was the method described and appropriate for assessing conceptually and theoretically sound hypothesized relationships?

    REFERENCE: Testing attachment, section 2b1.

    ***TIPS**: For example: correlation of the performance measure score on this measure and other performance measures; differences in performance scores between groups known to differ on quality; other accepted method with description of how it assesses validity of the performance score*

    ☒Yes (go to Question #20)

    ☐No (please explain below, then go to Question #20 and rate as INSUFFICIENT)

20. **RATING (measure score)** - Based on the measure score results (significance, strength) and scope of testing (number of measured entities and representativeness) and analysis of potential threats, what is the level of certainty or confidence that the performance measure scores are a valid indicator of quality?

    ☐High (go to Question #21)

    ☒Moderate (go to Question #21)

    ☐Low (please explain below then go to Question #21)

    ☐Insufficient (go to Question #21)

21. Was validity testing conducted with patient-level data elements?

    REFERENCE: Testing attachment, section 2b1.

    ***TIPS**: Prior validity studies of the same data elements may be submitted*

    ☒Yes (go to Question #22)

    ☐No (if there is score-level testing that you rated something other than INSUFFICIENT in Question #20, skip questions #22-25, and go to Question #26 (OVERALL VALIDITY); otherwise, skip questions #22-23 and go to Question #24)

22. Was the method described and appropriate for assessing the accuracy of ALL critical data elements? *NOTE that data element validation from the literature is acceptable.*

    ***REFERENCE:** Testing attachment, section 2b1.*

    ***TIPS**: For example: Data validity/accuracy as compared to authoritative source- sensitivity, specificity, PPV, NPV; other accepted method with description of how it assesses validity of the data elements.*

Answer No if: only assessed percent agreement; did not assess separately for all data elements (at least numerator, denominator, exclusions)

☐Yes (go to Question #23)

☒No (please explain below, then go to Question #23 and rate as INSUFFICIENT)

The developer only assessed by overall percent agreement. NQF guidance states that all critical elements must be assessed separately, which did not occur.

23. **RATING (data element)** - Based on the data element testing results (significance, strength) and scope of testing (number and representativeness of patients and entities) and analysis of potential threats, what is the level of certainty or confidence that the data used in the measure are valid?

    ☐Moderate (skip Questions #24-25 and go to Question #26)

    ☐Low (please explain below, skip Questions #24-25 and go to Question #26)

    ☒Insufficient (go to Question #24 only if no other empirical validation was conducted OR if the measure has not been previously endorsed; otherwise, skip Questions #24-25 and go to Question #26)

24. Was face validity systematically assessed by recognized experts to determine agreement on whether the computed performance measure score from the measure as specified can be used to distinguish good and poor quality?

    NOTE: If appropriate empirical testing has been conducted, then evaluation of face validity is not necessary; you should skip this question and Question 25, and answer Question #26 based on your answers to Questions #20 and/or #23]

    REFERENCE: Testing attachment, section 2b1.

    ***TIPS**: Answer no if: focused on data element accuracy/availability/feasibility/other topics; the degree of consensus and any areas of disagreement not provided/discussed.*

    ☐Yes (go to Question #25)

    ☐No (please explain below, skip question #25, go to Question #26 (OVERALL VALIDITY) and rate as INSUFFICIENT)

25. **RATING (face validity)** - Do the face validity testing results indicate substantial agreement that the performance measure score from the measure as specified can be used to distinguish quality AND potential threats to validity are not a problem, OR are adequately addressed so results are not biased?

    REFERENCE: Testing attachment, section 2b1.

    ***TIPS**: Face validity is no longer accepted for maintenance measures unless there is justification for why empirical validation is not possible and you agree with that justification.*

    ☐Yes (if a NEW measure, go to Question #26 (OVERALL VALIDITY) and rate as MODERATE)

    ☐ Yes (if a MAINTENANCE measure, do you agree with the justification for not conducting empirical testing? If no, go to Question #26 (OVERALL VALIDITY) and rate as INSUFFICIENT; otherwise, rate Question #26 as MODERATE)

    ☐No (please explain below, go to Question #26 (OVERALL VALIDITY) and rate AS LOW)

OVERALL VALIDITY RATING

26. **OVERALL RATING OF VALIDITY** taking into account the results and scope of all testing and analysis of potential threats.

    ☐High (NOTE: Can be HIGH only if score-level testing has been conducted)

    ☒Moderate (NOTE: Moderate is the highest eligible rating if score-level testing has NOT been conducted)

    ☐Low (please explain below) [NOTE: Should rate LOW if you believe that there are threats to validity and/or

threats to validity were not assessed]

☐Insufficient (if insufficient, please explain below) [NOTE: For instrument-based measures and some composite measures, testing at both the score level and the data element level is required; if not conducted, should rate as INSUFFICIENT—please check with NQF staff if you have questions.]

FOR COMPOSITE MEASURES ONLY: Empirical analyses to support composite construction

27. What is the level of certainty or confidence that the empirical analysis demonstrates that the component measures add value to the composite and that the aggregation and weighting rules are consistent with the quality construct?

REFERENCE: Testing attachment, section 2c

***TIPS**: Consider the following: Do the component measures fit the quality construct? Are the objectives of parsimony and simplicity achieved while supporting the quality construct?*

☐High

☐Moderate

☐Low (please explain below)

☐Insufficient (please explain below)

Committee Pre-evaluation Comments: Criteria 2: Scientific Acceptability of Measure Properties (including all 2a, 2b, and 2c)

2a1. Specifications:

- The specification of the denominator exclusions of second PCIs within 30 days is a bit confusing. Although it may not have a large impact, it appears that a PCI performed 20 days after an index PCI, followed by a death 29 days after that would not be attributed. Also, it is not clear what is meant by "additional PCI procedures within 30 days of the death" in the following: "Subsequent PCIs within 30-days. The 30-day outcome period for patients with more than one PCI may overlap. In order to avoid attributing the same death to more than one PCI (i.e. double counting a single patient death), additional PCI procedures within 30 days of the death are not counted as new index procedures.
- All data elements are clearly defined.
- Data elements appear to be clearly defined.
- Because of the high use of this measure, the reliability specifications are quite high. I believe that the codes are appropriate and easily identified for this measure says.
- The denominator quality assurance is uncertain. Does the developer ensure that all PCI procedures are reported that result in adverse outcome? Does the quality assurance process adequately ascertain that all hospital PCI procedures are indeed included in the hospital's final report?
- Specifications of variables clear for this measure (variables like shock would be less clear).
- Data elements are will defined as are the numerator and denominator.

2a2. Reliability testing:

- No concerns
- Reliability was judged to be moderate on last review, but I see no details in this section, so cannot evaluate the measure's reliability.
- I have no concerns on reliability.
- As above concerns about the data quality in the denominator, as the numerator appears more robust"
- Reliability of individual variables is high (>90%). Reliability of model estimates fair (~0.255).
- Test-retest reliability testing results are reasonable.

2b1. Validity testing & 2b4-7. Threats to Validity (Statistically Significant Differences, Multiple Data Sources, Missing Data):

- The data used to RSMRs is collected as part of routine care, so missing data is not an issue.
- Timeliness of the data collection appears problematic, undermining the currency of its validity status.
- The only concern is that this measure was not evaluated by a Scientific Methods panel, otherwise the validity if clearly there especially face validity.
- Missing data is key for the validity of this measure:
    1. non Medicare patients are not included, yet the measure title may be construed as reporting hospital mortality for ALL patients
    2. Denominator completeness is not assured
- 5X mortality differences meaningful for a generally low mortality procedure. Missing data low
- Data element testing is reasonable. I did note that on their testing of exclusions, there may be an error reported for the percent of deaths reported in 2010 (15.33%) compared to 1.86% in 2011 and 1.79% in 2010-2011. The same is true of the number of deaths reported in 2010 (23,699) vs. 2,315 in 2011, and 4,569 in 2010-2011. The 2010 seems to be off by a factor of 10. Developers should review this. C-statistic of model is 0.821. Not all critical data elements were tested.

2b2-3. Other Threats to Validity (Exclusions, Risk Adjustment):

- Exclusions appear to be appropriate
- Unclear that risk adjustment is acceptable.
- The risk adjustment does not take into account the elements of greater importance such as social determinant to health, compliance, management of other associated risk factors.
- Exclusion of non-Medicare population. No risk adjustment for race
- No inclusion of social factors. Distribution of hospital performance similar for hospitals caring for low SES patients at higher frequencies. More clinical data than claims measures should capture extent to which lower SES patients present "sicker"
- This is not adjusted for social risk factors. Overall missing values was < 1%, but 3 variables had significant missing values: BMI, LFEF, and GFR.

Criterion 3. Feasibility

Maintenance measures – no change in emphasis – implementation issues may be more prominent

3. Feasibility is the extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

- The developer stated that for clinical measures, the required data elements are routinely generated/collected during provision of care (e.g., blood pressure, lab value, diagnosis, medication order, depression score). Also, the data is abstracted from a record by another individual than the individual who obtained the original information (e.g., chart abstraction for quality measure/registry). However, the measure is obtained from an administrative database such as the National Death Index (NDI).
- All data elements in the electronic health records are in defined fields.
- The implementation of this measure requires patient data being matched to external data source to determine the outcome or death 30 days after PCI, which resulted in implementation challenges. These challenges are summarized below:
    1. Data availability: ACC could not implement this measure using CMS data and could not identify a mechanism to use the CMS data for public reporting. As a result of this, the developer had to modify

their implementation strategy, rework their models, and matched National Cardiovascular Data Registry (NCDR) records to CDC NDI data.

2. Patient Confidentiality: CDC NDI required direct patient identifiers to meet the minimum criteria for matching. Based on the 2017 Q2 CathPCI data, roughly 15% of the submitted NCDR sites did not submit direct patient identifiers to the registry; therefore ineligible for NDI matching and could not participate in this measure.
3. Data cost: Use of the CDC NDI data as a source of vital status have a substantial cost. For instance, retrieving the vital status on three years of PCI patients resulted in a cost of approximately $100,000.
4. Data Timeliness: CDC NDI is released yearly and about one year after the calendar year of death in addition to the processing time of the matching process, and the most contemporary data available over 18 months old.

- There are fees and licensing requirements for use of this measure in the ACC's NCDR program.

Questions for the Committee:

- *Are the required data elements routinely generated and used during care delivery?*
- *Are the required data elements available in electronic form, e.g., EHR or other electronic sources?*
- *Do you have any concerns about the feasibility of the measure since the developers mentioned implementation challenges?*

Preliminary rating for feasibility: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee Pre-evaluation Comments: Criteria 3: Feasibility

3. Feasibility

- Low – for the reasons they report:
  - Patient Confidentiality: CDC NDI required direct patient identifiers to meet the minimum criteria for matching. Based on the 2017 Q2 CathPCI data, roughly 15% of the submitted NCDR sites did not submit direct patient identifiers to the registry; therefore ineligible for NDI matching and could not participate in this measure.
  - Data cost: Use of the CDC NDI data as a source of vital status have a substantial cost. For instance, retrieving the vital status on three years of PCI patients resulted in a cost of approximately $100,000.
  - Data Timeliness: CDC NDI is released yearly and about one year after the calendar year of death in addition to the processing time of the matching process, and the most contemporary data available over 18 months old. There are fees and licensing requirements for use of this measure in the ACC's NCDR program.
- Feasibility is moderate. Scanning the NDI is expensive. Some centers may not submit all cases.
- Apparent serious problems with feasibility (data timeliness and validity).
- This is very feasible.
- It is unclear if Medicare claims data are cross checked with the reporting data to NCDR from hospitals (i.e., is there quality control to ascertain that all procedures that Medicare received claims for are actually included in NCDR).
- NDI data not generated during care delivery, but is available for NCDR and CMS.
- Noted factors that impact feasibility include the use of PII, cost of matching to National Death Index, and time lag to match with the National Death Index.

Criterion 4: Usability and Use

Maintenance measures – increased emphasis – much greater focus on measure use and usefulness, including both impact/improvement and unintended consequences

4a. Use (4a1. Accountability and Transparency; 4a2. Feedback on measure)

4a. Use evaluate the extent to which audiences (e.g., consumers, purchasers, providers, policymakers) use or could use performance results for both accountability and performance improvement activities.

4a.1. Accountability and Transparency. Performance results are used in at least one accountability application within three years after initial endorsement and are publicly reported within six years after initial endorsement (or the data on performance results are available). If not in use at the time of initial endorsement, then a credible plan for implementation within the specified timeframes is provided.

Current uses of the measure

Publicly reported? ☐ **Yes** ☒ **No**
Current use in an accountability program? ☐ **Yes** ☒ **No** ☐ **UNCLEAR**
OR
Planned use in an accountability program? ☐ **Yes** ☒ **No**

Accountability program details: The measure is designed for use in public reporting, but ACC held off on public reporting since because they are in the process of updating the CathPCI registry to version 5. The developer is also in the process of harmonizing #0536 and #0133 (in-hospital mortality). Additionally, ACC was in the process of transferring measure stewardship from CMS during the last re-endorsement in 2014. Therefore, ACC previously had limited control over the public use of the measure until it resumed primary stewardship of the measure less than 3 years ago. ACC plans to include this measure in NCDR's public reporting program in the future.

4a.2. Feedback on the measure by those being measured or others. Three criteria demonstrate feedback: 1) those being measured have been given performance results or data, as well as assistance with interpreting the measure results and data; 2) those being measured and other users have been given an opportunity to provide feedback on the measure performance or implementation; 3) this feedback has been considered when changes are incorporated into the measure

Feedback on the measure by those being measured or others: The developer stated that performance results are distributed to all CathPCI registry participants in the quarterly benchmark reports that provides a detailed analysis of an institution's individual performance in comparison to the entire registry population from participating hospitals across the nation. Additionally, when this measure was first implemented in the CathPCI registry in quarter 3 2017, registry participants expressed to be very interested in the measure.

Additional Feedback: The developer stated that feedback is usually obtained through monthly registry site manager calls, ad hoc phone calls tracked with salesforce software, and during registry/specific breakout sessions at the NCDR's annual meeting. The Registry Standing Committee members may provide feedback during regularly scheduled calls as well. The developer states that the majority of the data elements are in structured format within the patient's electronic health record and could be attained without undue burden within the hospital, which minimizes time and cost.

***Questions for the Committee**:*

- *How have (or can) the performance results be used to further the goal of high-quality, efficient healthcare?*
- *How has the measure been vetted in real-world settings by those being measured or others?*

Preliminary rating for Use: ☒ **Pass** ☐ **No Pass**

4b. Usability (4a1. Improvement; 4a2. Benefits of measure)

4b. Usability evaluate the extent to which audiences (e.g., consumers, purchasers, providers, policymakers) use or could use performance results for both accountability and performance improvement activities.

4b.1 Improvement. Progress toward achieving the goal of high-quality, efficient healthcare for individuals or populations is demonstrated.

Improvement results: The developers are unable to comment on or draw conclusions from the risk adjusted performance trends over time because of the differences in the cohorts of data analyzed (CMS versus NDI).

4b2. Benefits vs. harms. Benefits of the performance measure in facilitating progress toward achieving high-quality, efficient healthcare for individuals or populations outweigh evidence of unintended negative consequences to individuals or populations (if such evidence exists).

Unexpected findings (positive or negative) during implementation: Implementation challenges were outlined in the feasibility section as the availability, cost, and timeliness of data and concerns about patient confidentiality.

Potential harms: The developer stated studies suggest that public reporting of the outcomes of cardiovascular procedures may have unintended consequences but determining the underlying causes and appropriateness of these differences is not possible at this time. The developer noted concerns that physicians in states that publicly report PCI outcomes would either refer high risk cases to states without public reporting or avoid reporting such cases. This measure has not undergone public reporting to date, thus the unintended consequences are speculative.

Additional Feedback: None

***Questions for the Committee**:*

- *How can the performance results be used to further the goal of high-quality, efficient healthcare?*
- *Do the benefits of the measure outweigh any potential unintended consequences?*

Preliminary rating for Usability and use: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Rationale:

Committee Pre-evaluation Comments: Criteria 4: Usability and Use

4a. Use

- The measure is not yet reported because stewardship is being transferred from CMS to ACCF. ACCF declares that they plan to publicly report.
- Remaining issues with reporting problems, apparently.
- This is a measure that is working to develop wide spread adoption--but it is not there yet.
- Not publicly reported as yet.
- Recently included results in NCDR reports.
- Not currently publically reported, but there is a plan for this. This measure is in the process of transitioning from CMS stewardship to ACC stewardship.

4b. Usability

- There is some theoretical consideration that cardiologists will not operate on tough cases or will transfer them to states that do not publicly report. This is purely speculative.
- Unclear regarding unintended consequences and balance of benefits with potential harm.
- This measure is still rather coarse and requires the providers and the consumers to tease individual risk factors to modify the interpretation of the data.

- Potential harms are due to the measure design as it measures mortality in the Medicare population which may or may not be different from the mortality in the whole population, Public perception may be moved in one or another direction without specifically stressing that this measure only measures Medicare population mortality.
- Less likely to cause harm than parallel STEMI/shock measure because pulls out the sickest patients. Also clinical variables should in theory produce better risk adjustment. ROC 0.82. But will need to monitor for unintended consequences anyway.
- Some concerns about public reporting are mentioned - avoidance of challenging cases. Unintended consequences are speculative.

Criterion 5: Related and Competing Measures

Related or competing measures

- NQF #0229 : Hospital 30-day, all-cause, risk-standardized mortality rate (RSMR) following heart failure (HF) hospitalization
- NQF #0230 : Hospital 30-day, all-cause, risk-standardized mortality rate (RSMR) following acute myocardial infarction (AMI) hospitalization for patients 18 and older
- NQF #0536 : 30-day all-cause risk-standardized mortality rate following Percutaneous Coronary Intervention (PCI) for patients with ST segment elevation myocardial infarction (STEMI) or cardiogenic shock

Harmonization

- The developer stated that this measure is harmonized to the extent possible to the related and competing measures.

Committee Pre-evaluation Comments: Criterion 5: Related and Competing Measures

- None

Public and Member Comments

Comments and Member Support/Non-Support Submitted as of: June 12, 2018

No comments have been submitted as of this date.

1. Evidence and Performance Gap – Importance to Measure and Report

Extent to which the specific measure focus is evidence-based, important to making significant gains in healthcare quality, and improving health outcomes for a specific high-priority (high-impact) aspect of healthcare where there is variation in or overall less-than-optimal performance. ***Measures must be judged to meet all sub criteria to pass this criterion and be evaluated against the remaining criteria.***

1a. Evidence to Support the Measure Focus – See attached Evidence Submission Form

0535_NQF_evidence_attachment_Sep2017_2.26.18.docx

1a.1 For Maintenance of Endorsement: Is there new evidence about the measure since the last update/submission? Do not remove any existing information. If there have been any changes to evidence, the Committee will consider the new evidence. Please use the most current version of the evidence attachment (v7.1). Please use red font to indicate updated evidence.

No

1a Evidence (subcriterion 1a)

Measure Number (*if previously endorsed*)**:** 0535

Measure Title: 30-day all-cause risk-standardized mortality rate following Percutaneous Coronary Intervention (PCI) for patients without ST segment elevation myocardial infarction (STEMI) and without cardiogenic shock

IF the measure is a component in a composite performance measure, provide the title of the Composite Measure here:

Date of Submission: 11/8/2017

Instructions

- *Complete 1a.1 and 1a.2 for all measures. If instrument-based measure, complete 1a.3.*
- *Complete **EITHER 1a.2, 1a.3 or 1a.4** as applicable for the type of measure and evidence.*
- *For composite performance measures:*
  - *A separate evidence form is required for each component measure unless several components were studied together.*
  - *If a component measure is submitted as an individual performance measure, attach the evidence form to the individual measure submission.*
- All information needed to demonstrate meeting the evidence subcriterion (1a) must be in this form. An appendix of *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.
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.

Note: The information provided in this form is intended to aid the Standing Committee and other stakeholders in understanding to what degree the evidence for this measure meets NQF's evaluation criteria.

1a. Evidence to Support the Measure Focus

The measure focus is evidence-based, demonstrated as follows:

- Outcome: [3] Empirical data demonstrate a relationship between the outcome and at least one healthcare structure, process, intervention, or service. If not available, wide variation in performance can be used as evidence, assuming the data are from a robust number of providers and results are not subject to systematic bias.

- Intermediate clinical outcome: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured intermediate clinical outcome leads to a desired health outcome.
- Process: [5] a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured process leads to a desired health outcome.
- Structure: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured structure leads to a desired health outcome.
- Efficiency: [6] evidence not required for the resource use component.
- For measures derived from patient reports, evidence should demonstrate that the target population values the measured outcome, process, or structure and finds it meaningful.
- Process measures incorporating Appropriate Use Criteria: See NQF's guidance for evidence for measures, in general; guidance for measures specifically based on clinical practice guidelines apply as well.

Notes

3. Generally, rare event outcomes do not provide adequate information for improvement or discrimination; however, serious reportable events that are compared to zero are appropriate outcomes for public reporting and quality improvement.

4. The preferred systems for grading the evidence are the Grading of Recommendations, Assessment, Development and Evaluation (GRADE) guidelines and/or modified GRADE.

5. Clinical care processes typically include multiple steps: assess → identify problem/potential problem → choose/plan intervention (with patient input) → provide intervention → evaluate impact on health status. If the measure focus is one step in such a multistep process, the step with the strongest evidence for the link to the desired outcome should be selected as the focus of measurement. Note: A measure focused only on collecting PROM data is not a PRO-PM.

6. Measures of efficiency combine the concepts of resource use and quality (see NQF's Measurement Framework: Evaluating Efficiency Across Episodes of Care; AQA Principles of Efficiency Measures).

1a.1.This is a measure of: (*should be consistent with type of measure entered in De.1*)

Outcome

☒ Outcome: 30-day mortality

☐Patient-reported outcome (PRO):

PROs include HRQoL/functional status, symptom/symptom burden, experience with care, health-related behaviors. (*A PRO-based performance measure is not a survey instrument. Data may be collected using a survey instrument to construct a PRO measure.)*

☐ Intermediate clinical outcome (*e.g., lab value*):

☐ Process:

☐ Appropriate use measure:

☐ Structure:

☐ Composite:

1a.2 LOGIC MODEL Diagram or briefly describe the steps between the healthcare structures and processes (e.g., interventions, or services) and the patient's health outcome(s). The relationships in the diagram should be easily understood by general, non-technical audiences. Indicate the structure, process or outcome being measured.

The goal of this measure is to reduce PCI 30-day mortality rates to best-in-class. Measurement of patient outcomes, including mortality, allows for a broad view of quality of care that encompasses more than what can be captured by individual process-of-care measures. As described below, mortality is likely to be influenced by a broad range of clinical activities such as the prevention of complications and the provision of evidenced-based care.

1a.3 Value and Meaningfulness: IF this measure is derived from patient report, provide evidence that the target population values the measured ***outcome, process, or structure*** and finds it meaningful. (Describe how and from whom their input was obtained.)

Evidence that the outcome measure has been influenced by one or more clinical interventions:

Numerous studies have demonstrated the efficacy of interventions designed to improve patient outcomes following PCI. These include pharmacologic interventions such as the use of glycoprotein 2b/3a inhibitors, direct thrombin inhibitors, and pre-procedural thienopyridines such as clopidogrel and prasugrel, as well as advances in device technology such as use of stents,, thrombectomy for acute lesions with high thrombus burden, and distal embolic protection for PCI of degenerated saphenous vein grafts. Of note, the majority of these interventions have been shown to reduce endpoints other than mortality, most commonly rates of periprocedural MI, major bleeding, and target vessel revascularization for in-stent restenosis. Although few individual interventions have been shown to reduce mortality, they may collectively exert a favorable impact on hospital PCI mortality rates when implemented in a coordinated fashion.

There is a growing body of evidence that quality improvement efforts can improve outcomes of PCI patients, including survival. Rihal and colleagues examined patient outcomes before and after initiation of a program of continuous quality improvement (CQI) and found a significantly lower in-hospital mortality following PCI despite significant increases in the risk profile of PCI patients. Similar improvements were identified in studies of CQI by Brush et al and Moscucci et al, and improvements in survival were associated with greater adherence to evidence-based practices including preprocedural clopidogrel, use of glycoprotein 2b/3a inhibitors, and volume of iodinated contrast. The observational nature of these studies precludes drawing definitive conclusions, but they strongly suggest a

mechanism through which public reporting of hospital PCI outcomes could promote improvements in the care of PCI patients.

References:

Brush JE, Balakrishnan SA, Brough J, Hartman C, Hines G, Liverman DP, Parker JP, Rich J, Tindall N. (2006). "Implementation of a continuous quality improvement program for percutaneous coronary intervention and cardiac surgery at a large community hospital." Am Heart J 152 (2):379-85 16875926 (P,S,E,B).

Krumholz HM, Brindis RG, et al. (2006). "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 113(3): 456-62.

Moscucci M, Kline Rogers E, Montoye C, Smith DE, Share D, O'Donnell M, Maxwell-Eward A, Meengs WL, De Franco AC, Patel K, McNamara R, McGinnity JG, Jani SM, Khanal S, Eagle KA. (2006). "Association of a Continuous Quality Improvement Initiative With Practice and Outcome Variations of Contemporary Percutaneous Coronary Interventions." Circulation. 113:814-822.

Rihal C, Kamath C, Holmes D, et al. (2006). "Economic and clinical outcomes of a physician-led continuous quality improvement intervention in the delivery of percutaneous coronary intervention." Am J Manag Care 12:445-452.

****RESPOND TO ONLY ONE SECTION BELOW -EITHER 1a.2, 1a.3 or 1a.4) ****

1a.2 FOR OUTCOME MEASURES including PATIENT REPORTED OUTCOMES - Provide empirical data demonstrating the relationship between the outcome (or PRO) to at least one healthcare structure, process, intervention, or service.

See evidence/literature described above in 1a3.

1a.3. SYSTEMATIC REVIEW(SR) OF THE EVIDENCE (for INTERMEDIATE OUTCOME, PROCESS, OR STRUCTURE PERFORMANCE MEASURES, INCLUDING THOSE THAT ARE INSTRUMENT-BASED) If the evidence is not based on a systematic review go to section 1a.4) If you wish to include more than one systematic review, add additional tables.

What is the source of the systematic review of the body of evidence that supports the performance measure? A systematic review is a scientific investigation that focuses on a specific question and uses explicit, prespecified scientific methods to identify, select, assess, and summarize the findings of similar but separate studies. It may include a quantitative synthesis (meta-analysis), depending on the available data. (IOM)

☐ Clinical Practice Guideline recommendation (with evidence review)

☐ US Preventive Services Task Force Recommendation

☐ Other systematic review and grading of the body of evidence (*e.g., Cochrane Collaboration, AHRQ Evidence Practice Center*)

☐ Other

Source of Systematic Review: • **Title** • **Author** • **Date** • **Citation, including page number** • **URL**	
Quote the guideline or recommendation verbatim about the process, structure or intermediate outcome being measured. If not a guideline, summarize the conclusions from the SR.	
Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	
Provide all other grades and definitions from the evidence grading system	
Grade assigned to the **recommendation** with definition of the grade	
Provide all other grades and definitions from the recommendation grading system	
Body of evidence: • Quantity – how many studies? • Quality – what type of studies?	
Estimates of benefit and consistency across studies	
What harms were identified?	
Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	

_____

1a.4 OTHER SOURCE OF EVIDENCE

If source of evidence is NOT from a clinical practice guideline, USPSTF, or systematic review, please describe the evidence on which you are basing the performance measure.

1a.4.1 Briefly SYNTHESIZE the evidence that supports the measure. A list of references without a summary is not acceptable.

1a.4.2 What process was used to identify the evidence?

1a.4.3. Provide the citation(s) for the evidence.

1b. Performance Gap

Demonstration of quality problems and opportunity for improvement, i.e., data demonstrating:

- considerable variation, or overall less-than-optimal performance, in the quality of care across providers; and/or
- Disparities in care across population groups.

1b.1. Briefly explain the rationale for this measure *(e.g., how the measure will improve the quality of care, the benefits or improvements in quality envisioned by use of this measure)*

If a COMPOSITE (e.g., combination of component measure scores, all-or-none, any-or-none), SKIP this question and answer the composite questions.

This measure will describe hospital-level mortality rates following PCI in patients without STEMI and without cardiogenic shock, with the overriding goal to reduce 30-day mortality rates to best-in-class. The expectation is that providing this information to hospitals, coupled with public reporting of hospitals' results, will drive internal hospital quality improvement efforts to focus efforts on reducing PCI mortality. Of note, the measure includes not only in-hospital deaths, but also deaths occurring after hospital discharge. This perspective may motivate hospitals to look for opportunities not only within the organization, but to better coordinate the transition of care from the inpatient to the outpatient arena.

1b.2. Provide performance scores on the measure as specified (current and over time) at the specified level of analysis. *(This is required for maintenance of endorsement. Include mean, std dev, min, max, interquartile range, scores by decile. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities include.) This information also will be used to address the sub-criterion on improvement (4b1) under Usability and Use.*

Please see 1b.2. answer in supplemental attachment entitled " 0535F_Main Submission Form Supplement_04.09.18" and below.

The study cohort for the validation of this measure includes NCDR CathPCI data linked with National Death Index (NDI) data to ascertain the specifications for 30-day RSMRs for all payers and all ages (>18 years). Using the previously endorsed measure (note: there have been no changes to the measure specifications), we analyzed variation in 30-day RSMRs among the hospitals in this linked dataset for a three-year period, from December 2011 to December 2014. We excluded two months of observation due to missing data during our sampling frame (October 2012 and November 2012). There were 1,127,423 admissions to 1,365 hospitals in the combined three-year sample. RSMRs varied among hospitals, with a mean of 1.07%, a standard deviation of 0.30%, and a range of 0.51% to 2.70%. The interquartile range was 0.87% to 1.24%. The range of performance is as follows:

Percentile of RSMR	Mean RSMR
100% Max	0.0270
99%	0.0211
95%	0.0163
90%	0.0147
75% Q3	0.0124
50% Median	0.0104
25% Q1	0.0087
10%	0.0073
5%	0.0067

Percentile of RSMR	Mean RSMR
1%	0.0057
0% Min	0.0051

Note: The measure is oriented so a lower rate equals better performance. For the purposes of interpreting the above tables, interpret the 0% Min as hospitals with the lowest mortality rate and 100% max as hospitals with the highest mortality rate (i.e. percentiles correspond with mortality rates not percentiles of performance). For example, among all hospitals, the top 1% of hospitals have a mortality rate of 0.57% vs hospitals ranked in the 99th percentile have a 2.1% mortality rate.

Table 1: Distribution of hospital 30-day RSMR for NSTEMI/No Shock, 2011-2014

Description and Percentile	Mean RSMR
N	1127423
Mean	0.0107
Std Deviation	0.0030
100% Max	0.0270
99%	0.0211
95%	0.0163
90%	0.0147
75% Q3	0.0124
50% Median	0.0104
25% Q1	0.0087
10%	0.0073
5%	0.0067
1%	0.0057
0% Min	0.0051

Figure 1: Histogram of hospital 30-day RSMR for NSTEMI/No Shock, 2011-2014

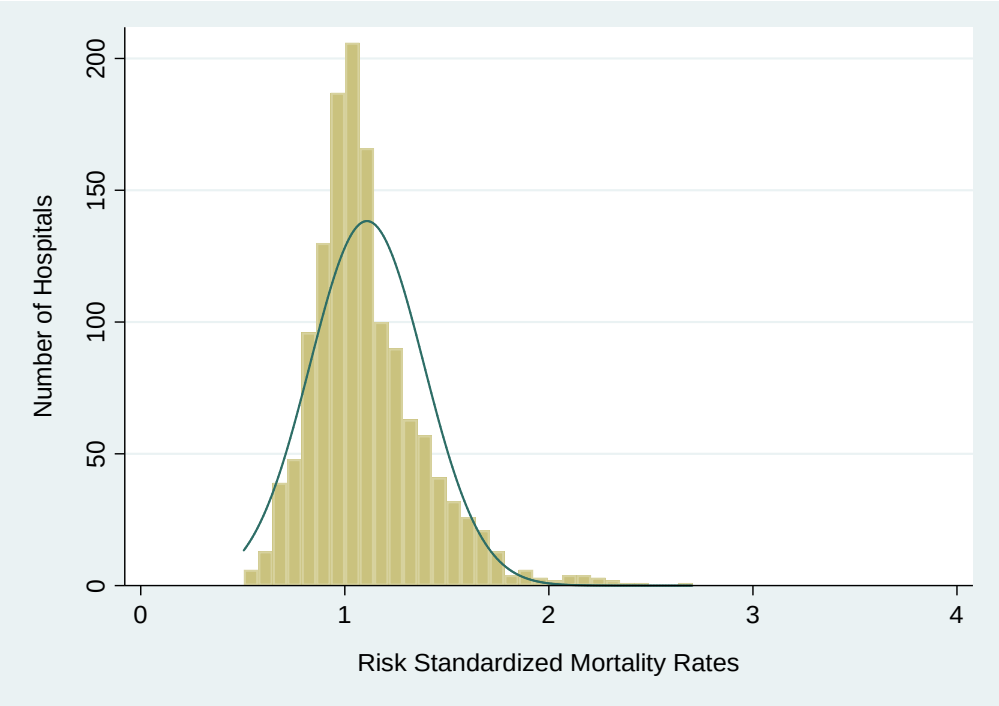
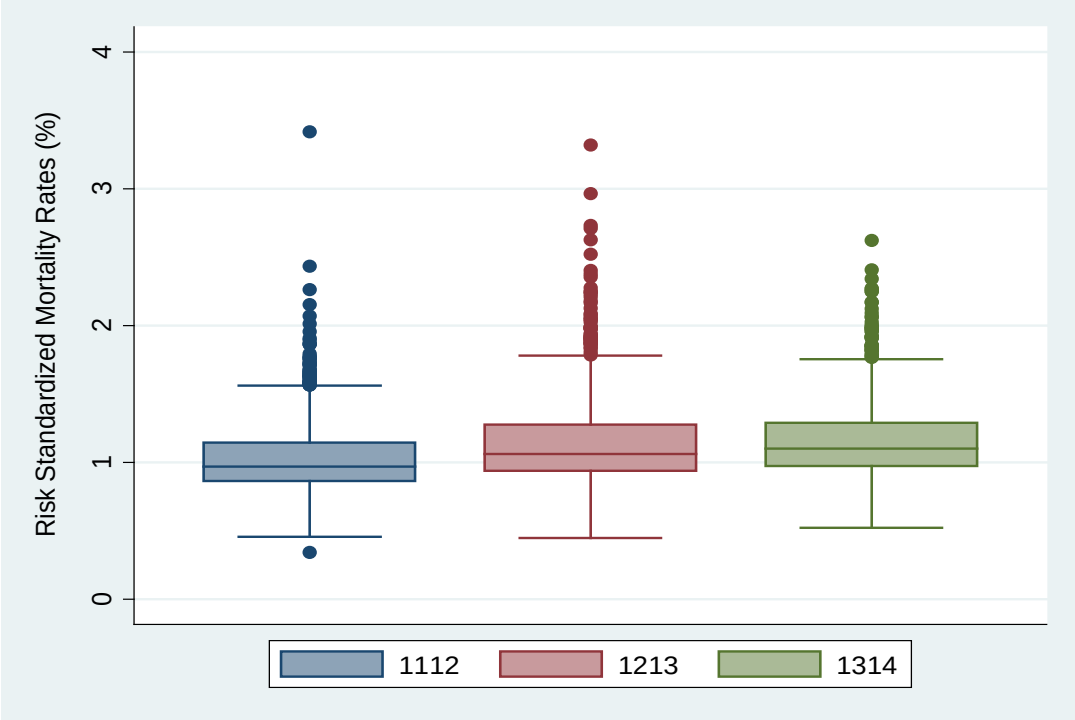


Table 2: Distribution of hospital 30-day RSMR for NSTEMI/No Shock by year

Description and Percentile	Mean RSMR by Year		
	2011-2012	2012-2013	2013-2014
N	350941	345929	430553
Mean	0.0100	0.0112	0.0113
Std Deviation	0.0029	0.0034	0.0030
100% Max	0.0342	0.0332	0.0262
99%	0.0178	0.0221	0.0206
95%	0.0154	0.0175	0.0168
90%	0.0137	0.0154	0.0151
75% Q3	0.0115	0.0129	0.0129
50% Median	0.0095	0.0106	0.0108
25% Q1	0.0082	0.0089	0.0091
10%	0.0068	0.0076	0.0079
5%	0.0061	0.0069	0.0072
1%	0.0049	0.0060	0.0061
0% Min	0.0034	0.0045	0.0052

Figure 2: Boxplot of hospital 30-day RSMR for STEMI/Shock, by year



1b.3. If no or limited performance data on the measure as specified is reported in 1b2, then provide a summary of data from the literature that indicates opportunity for improvement or overall less than optimal performance on the specific focus of measurement.

Inpatient mortality is the indicator that has been most widely used to evaluate the quality of cardiac procedures and is arguably the most important adverse outcome measure. The ACC summarized the experience of the NCDR CathPCI Registry from 1998-2000 and found that in-hospital mortality occurred in 1,422 of 100,253 PCI procedures (1.4%) (Shaw, Anderson et al. 2002). Mortality was higher in patients with acute myocardial infarction (4.9%) or cardiogenic shock (27.2%). In the present era, mortality rates for PCI in large series from experienced operators varied across hospitals (Carrozza, Cutlip et al. 2008). Prior studies have demonstrated significant variability in in-hospital PCI mortality across age groups, gender, geographic regions, socioeconomic status, and by hospital volume (Mukherjee, Wainess et al. 2005). Although 12 states already report PCI outcomes, to date there has not been a unified national effort to publicly report PCI mortality.

Citations

Carrozza J, Cutlip D, Levin T. (2008). Periprocedural complications of percutaneous coronary intervention. UpToDate. B. Rose. Waltham, MA.

Mukherjee D, Wainess RM, et al. (2005). "Variation in outcomes after percutaneous coronary intervention in the United States and predictors of periprocedural mortality." Cardiology 103(3): 143-7.

Shaw RE, Anderson HV, et al. (2002). "Development of a risk adjustment mortality model using the American College of Cardiology-National Cardiovascular Data Registry (ACC-NCDR) experience: 1998-2000." J Am Coll Cardiol 39(7): 1104-12.

Rosamond W, Flegal K, Furie K, Go A, Greenlund K, Haase N, Hailpern SM, Ho M, Howard V, Kissela B, Kittner S, Lloyd-Jones D, McDermott M, Meigs J, Moy C, Nichol G, O'Donnell C, Roger V, Sorlie P, Steinberger J, Thom T, Wilson M, Hong Y. Heart Disease and Stroke Statistics_2008 Update: A Report From the American Heart Association Statistics Committee and Stroke Statistics Subcommittee and for the American Heart Association Statistics Committee and Stroke Statistics Subcommittee Circulation 2008;117;e25-e146; originally published online Dec 17, 2007; DOI: 10.1161/CIRCULATIONAHA.107.187998.

1b.4. Provide disparities data from the measure as specified (current and over time) by population group, e.g., by race/ethnicity, gender, age, insurance status, socioeconomic status, and/or disability. *(This is required for maintenance of endorsement. Describe the data source including number of measured entities; number of patients; dates of data; if a*

sample, characteristics of the entities included.) For measures that show high levels of performance, i.e., "topped out", disparities data may demonstrate an opportunity for improvement/gap in care for certain sub-populations. This information also will be used to address the sub-criterion on improvement (4b1) under Usability and Use.

Please see response to 1b.4 in the supplemental attachment entitled "0535F_Main Submission Form Supplement_04.09.18" and below.

We analyzed whether disparities in performance on this measure exist at the hospital-level by race and hospital safety net status.

To identify potential disparities by race, we examined the relationship between hospital-level RSMR and hospital proportion of non-White patients among all hospitals grouped by quartile of the proportion of non-White patients.

Analyses demonstrated that the median RSMR for hospitals with the highest quartile of non-White patients was 1.04% compared with 1.07% among hospitals with the lowest quartile of non-White patients. The distributions for the RSMRs overlapped, and many hospitals caring for the highest quartile of non-White patients performed well or better on this measure. In addition, in comparison to the registry mean RSMR of 1.07%, hospitals with the highest proportions of non-White patients do not have worse 30-day RSMRs in the CathPCI-NDI linked cohort.

Distribution of 30-day RSMR for NSTEMI/No Shock Stratified by Quartile of Non-White Patients

Description	RMSRs by Hospital Quartile of Non-White Patients			
	Q1	Q2	Q3	Q4
N	341	341	342	341
Mean	0.0112	0.0109	0.0112	0.0111
Std Deviation	0.0027	0.0027	0.0028	0.0030
100% Max	0.0270	0.0227	0.0232	0.0245
99%	0.0199	0.0183	0.0213	0.0216
95%	0.0162	0.0155	0.0163	0.0167
90%	0.0150	0.0146	0.0147	0.0147
75% Q3	0.0123	0.0124	0.0125	0.0125
50% Median	0.0107	0.0105	0.0107	0.0104
25% Q1	0.0096	0.0090	0.0093	0.0093
10%	0.0085	0.0078	0.0082	0.0081
5%	0.0079	0.0071	0.0076	0.0071
1%	0.0065	0.0060	0.0067	0.0059
0% Min	0.0060	0.0051	0.0058	0.0056

Similarly, to identify potential disparities related to socoioeconomic status (SES), we examined the relationship between RSMR and hospital safety net status. Safety net status was defined as government (public) hospitals or non-government hospitals with a caseload that is higher than the average of the Medicaid caseloads of hospitals within a given state plus one standard deviation of Medicaid caseload of hospitals within that state. We used the American Hospital Association data (2010) to calculate the Medicaid caseload and define hospital safety net status (Yes/No). Hospital safety net status was used as a marker of SES because safety net hospitals serve a low income and vulnerable patient population.

Analyses demonstrated that the median RSMR was 1.08% for safety net hospitals compared with 1.05% for non-safety net hospitals. The interquartile range for safety net hospitals was 0.95% to 1.31%, whereas among non-safety net

hospitals it was 0.93% to 1.23%. Overall, hospitals with a high proportion of vulnerable patients, as defined by safety net status, do not have substantially worse 30-day RSMRs in this cohort.

Consistent with NQF guidelines, this measure does not risk adjust for race or SES. Also, the results of the disparity data do not suggest the need to stratify by SDS.

Distribution of 30-day RSMR for NSTEMI/No Shock Stratified by Hospital Safety Net Status

Description	Safety Net Status	
	No	Yes
N	1025	205
Mean	0.0110	0.0117
Std Deviation	0.0027	0.0032
100% Max	0.0270	0.0237
99%	0.0197	0.0222
95%	0.0159	0.0174
90%	0.0145	0.0162
75% Q3	0.0123	0.0131
50% Median	0.0105	0.0108
25% Q1	0.0093	0.0095
10%	0.0080	0.0084
5%	0.0073	0.0078
1%	0.0060	0.0069
0% Min	0.0051	0.0063

1b.5. If no or limited data on disparities from the measure as specified is reported in 1b.4, then provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement. Include citations. Not necessary if performance data provided in 1b.4

N/A

2. Reliability and Validity—Scientific Acceptability of Measure Properties

Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. ***Measures must be judged to meet the sub criteria for both reliability and validity to pass this criterion and be evaluated against the remaining criteria.***

2a.1. Specifications The measure is well defined and precisely specified so it can be implemented consistently within and across organizations and allows for comparability. eMeasures should be specified in the Health Quality Measures Format (HQMF) and the Quality Data Model (QDM).

De.5. Subject/Topic Area *(check all the areas that apply):*

Cardiovascular, Cardiovascular : Coronary Artery Disease (PCI)

De.6. Non-Condition Specific*(check all the areas that apply):*

Care Coordination, Safety, Safety : Complications

De.7. Target Population Category *(Check all the populations for which the measure is specified and tested if any):*

Elderly, Populations at Risk

S.1. Measure-specific Web Page *(Provide a URL link to a web page specific for this measure that contains current detailed specifications including code lists, risk model details, and supplemental materials. Do not enter a URL linking to a home page or to general information.)*

https://www.ncdr.com/WebNCDR/docs/public-data-collection-documents/cathpci_v4_codersdictionary_4-4.pdf?sfvrsn=2

S.2a. If this is an eMeasure, HQMF specifications must be attached. Attach the zipped output from the eMeasure authoring tool (MAT) - if the MAT was not used, contact staff. (Use the specification fields in this online form for the plain-language description of the specifications)

This is not an eMeasure **Attachment:**

S.2b. Data Dictionary, Code Table, or Value Sets *(and risk model codes and coefficients when applicable) must be attached. (Excel or csv file in the suggested format preferred - if not, contact staff)*

Attachment **Attachment:** 0535F_PCI_Mortality_No_STEMI_3.15.18.xlsx

S.2c. Is this an instrument-based measure (i.e., data collected via instruments, surveys, tools, questionnaires, scales, etc.)? Attach copy of instrument if available.

No, this is not an instrument-based measure **Attachment:**

S.2d. Is this an instrument-based measure (i.e., data collected via instruments, surveys, tools, questionnaires, scales, etc.)? Attach copy of instrument if available.

Not an instrument-based measure

S.3.1. For maintenance of endorsement: Are there changes to the specifications since the last updates/submission. If yes, update the specifications for S1-2 and S4-22 and explain reasons for the changes in S3.2.

No

S.3.2. For maintenance of endorsement, please briefly describe any important changes to the measure specifications since last measure update and explain the reasons.

No changes were made to the measure specification since the last endorsement

S.4. Numerator Statement *(Brief, narrative description of the measure focus or what is being measured about the target population, i.e., cases from the target population with the target process, condition, event, or outcome) DO NOT include the rationale for the measure.*

IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm (S.14).

The outcome for this measure is all–cause death within 30 days following a PCI procedure in patients without STEMI and without cardiogenic shock at the time of the procedure.

S.5. Numerator Details *(All information required to identify and calculate the cases from the target population with the target process, condition, event, or outcome such as definitions, time period for data collection, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*

IF an OUTCOME MEASURE, describe how the observed outcome is identified/counted. Calculation of the risk-adjusted outcome should be described in the calculation algorithm (S.14).

Deaths can be identified using an external source of vital status, such as the Social Security Administration's Death Master File (DMF) or the Centers for Disease Control and Prevention's National Death Index (NDI). For the purpose of development and reassessment of the measure, we used a Medicare FFS population age 65 and over. We linked CathPCI registry with corresponding Medicare data and identified: a) in-hospital deaths using the discharge disposition indicator

in the Standard Analytic File (SAF) and identified) post-discharge deaths using the Enrollment Database (EDB). For the purpose of maintenance, the measure used a cohort of patients whose vital status was determined from the National Death Index. This data sample reflects a more comprehensive data set including a broader age range (>18 years) and an all-payer model compared to the Medicare data set (>65 years) used for initial measure testing.

S.6. Denominator Statement *(Brief, narrative description of the target population being measured)*

The target population for this measure includes inpatient and outpatient hospital stays with a PCI procedure for patients at least 18 years of age, without STEMI and without cardiogenic shock at the time of procedure.

S.7. Denominator Details *(All information required to identify and calculate the target population/denominator such as definitions, time period for data collection, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b.)*

IF an OUTCOME MEASURE, describe how the target population is identified. Calculation of the risk-adjusted outcome should be described in the calculation algorithm (S.14).

The time window can be specified from one or more years. This measure was developed with Medicare claims and CathPCI Registry data from one calendar year.

The measure cohort is patients undergoing PCI who do NOT have STEMI and do NOT have cardiogenic shock. STEMI or cardiogenic shock is defined as present in Version 4.4 of the CathPCI registry as follows:

Admissions with PCI are identified by field 5305 (PCI=yes);

STEMI or shock is identified by:

(1) Symptoms present on admission = ACS:STEMI (field 5000 = 6) with Time Period Symptom Onset to Admission within 24 hours (field 5005 = 5006, 5007, 5008) or Acute PCI = Yes (field 7035);

OR

(2) Cardiogenic shock = Yes (field 5060=1)

S.8. Denominator Exclusions *(Brief narrative description of exclusions from the target population)*

Hospital stays are excluded from the cohort if they meet any of the following criteria:

(1) PCIs that follow a prior PCI in the same admission (either at the same hospital or a PCI performed at another hospital prior to transfer).

This exclusion is applied in order to avoid assigning the death to two separate admissions.

(2) For patients with inconsistent or unknown vital status or other unreliable data (e.g. date of death precedes date of PCI);

(3) Subsequent PCIs within 30-days. The 30-day outcome period for patients with more than one PCI may overlap. In order to avoid attributing the same death to more than one PCI (i.e. double counting a single patient death), additional PCI procedures within 30 days of the death are not counted as new index procedures.

(4) PCIs for patients with more than 10 days between date of admission and date of PCI. Patients who have a PCI after having been in the hospital for a prolonged period of time are rare and represent a distinct population that likely has risk factors related to the hospitalization that are not well quantified in the registry.

S.9. Denominator Exclusion Details *(All information required to identify and calculate exclusions from the denominator such as definitions, time period for data collection, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b.)*

Excluded hospital stays are identified as follows:

(1) PCIs that follow a prior PCI in the same admission or occur during a transfer-in admission (PCI to PCI). For the purposes of development we used Medicare data to define transfers as two admissions that occur within 1 day of each other and identified patients in this cohort who had a PCI during both admissions. This can also be identified in the

registry data. (Note: For purposes of maintenance, we used CathPCI registry data to identify patients transferred in who had a prior PCI at the transferring hospital)

(2) Patients with inconsistent or unknown vital status or other unreliable data (e.g. date of death precedes date of PCI). The specific data fields will depend on the data source used.

(3) Not the first hospital stay with a PCI in the 30 days prior to a patient death. These stays are identified by procedure date in the CathPCI Registry and death date in the vital status data source.

(4) PCIs for patients with more than 10 days between date of admission and date of PCI. We determine length of stay by subtracting the admission date from the procedure date in the CathPCI Registry.

S.10. Stratification Information *(Provide all information required to stratify the measure results, if necessary, including the stratification variables, definitions, specific data collection items/responses, code/value sets, and the risk-model covariates and coefficients for the clinically-adjusted version of the measure when appropriate – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format with at S.2b.)*

Results of this measure will not be stratified.

S.11. Risk Adjustment Type (Select type. Provide specifications for risk stratification in measure testing attachment)

Statistical risk model

If other:

S.12. Type of score:

Rate/proportion

If other:

S.13. Interpretation of Score *(Classifies interpretation of score according to whether better quality is associated with a higher score, a lower score, a score falling within a defined interval, or a passing score)*

Better quality = Lower score

S.14. Calculation Algorithm/Measure Logic (*Diagram or describe the calculation of the measure score as an ordered sequence of steps including identifying the target population; exclusions; cases meeting the target process, condition, event, or outcome; time period for data, aggregating data; risk adjustment; etc.)*

The measure score is calculated based on the following steps:

1. Patient cohort is identified based on the inclusion and exclusion criteria (see questions S.7, S.8, S.9, S.10, S.11);
2. Data elements for risk adjustment are collected using the first collected value, as detailed below;
3. Outcome is ascertained from an outside data source, such as the Medicare Enrollment Database (see questions S.4, S.5, S.6)
4. Measure score is calculated with aggregated data across all included sites, as described below.

Risk-adjustment variables

The measure is adjusted for the variables listed below:

1. Age (10 year increments)
2. Body Mass Index (5 kg/m^2 increments)
3. History of congestive heart failure
5. History of cerebrovascular disease
6. History of peripheral vascular disease
7. History of chronic lung disease
8. Diabetes
9. Glomerular Filtration Rate (GFR) (derived)
10. Previous PCI
11. Heart Failure - current status

12. New York Hospital Association
13. Symptom onset
14. Ejection Fraction percent (EF)
15. PCI status
16. Highest risk lesion – coronary artery segment category
17. Highest risk lesion: Society for Cardiovascular Angiography and Interventions (SCAI)

Measure Score Calculation

The RSMR is calculated as the ratio of the number of "predicted" to the number of "expected" deaths, multiplied by the national unadjusted mortality rate. For each hospital, the predicted hospital outcome (the numerator) is the number of deaths within 30 days predicted on the basis of the hospital's performance with its observed case mix, and the "denominator" is the number of deaths expected on the basis of the nation's performance with that hospital's case mix. This approach is analogous to a ratio of "observed" to "expected" used in other types of statistical analyses. It conceptually allows for a comparison of a particular hospital's performance given its case mix to an average hospital's performance with the same case mix. Thus, a lower ratio indicates lower-than-expected mortality (better quality) and a higher ratio indicates higher-than-expected mortality (worse quality).

The predicted hospital outcome (the numerator) is calculated by regressing the risk factors and the hospital-specific intercept on the risk of mortality, multiplying the estimated regression coefficients by the patient characteristics in the hospital, transforming, then summing over all patients attributed to the hospital to get a value. The expected number of deaths (the denominator) is obtained by regressing the risk factors and a common intercept on the mortality outcome using all hospitals in our sample, multiplying the subsequent estimated regression coefficients by the patient characteristics observed in the hospital, transforming, and then summing over all patients in the hospital to get a value. To assess hospital performance in any reporting period, we re-estimate the model coefficients using the years of data in that period.

Please see attachments for more details on the calculation algorithm and the value sets for the risk-adjustment variables.

References:

Normand S-LT, Shahian DM. 2007. Statistical and Clinical Aspects of Hospital Outcomes Profiling. Stat Sci 22 (2): 206-226.

S.15. Sampling *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*

IF an instrument-based performance measure (e.g., PRO-PM), identify whether (and how) proxy responses are allowed.

N/A. This measure is not based on a sample or survey. Data from all hospitals and all PCI procedures would be included in the process of re-estimating model variables. For public reporting, minimum sample size has not been determined.

S.16. Survey/Patient-reported data *(If measure is based on a survey or instrument, provide instructions for data collection and guidance on minimum response rate.)*

Specify calculation of response rates to be reported with performance measure results.

N/A. This measure is not based on a sample or survey.

S.17. Data Source *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*

If other, please describe in S.18.

Claims, Other, Registry Data

S.18. Data Source or Collection Instrument *(Identify the specific data source/data collection instrument (e.g. name of database, clinical registry, collection instrument, etc., and describe how data are collected.)*

IF instrument-based, identify the specific instrument(s) and standard methods, modes, and languages of administration.

Data sources:

NCDR CatchPCI Registry

Vital Status Source:

National Death Index, Death Masterfile, Medicare enrollment database, or equivalent

S.19. Data Source or Collection Instrument *(available at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*

Available at measure-specific web page URL identified in S.1

S.20. Level of Analysis *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*

Facility, Other

S.21. Care Setting *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*

Inpatient/Hospital

If other:

S.22. <u>COMPOSITE Performance Measure</u> - Additional Specifications *(Use this section as needed for aggregation and weighting rules, or calculation of individual performance measures if not individually endorsed.)*

N/A

2. Validity – See attached Measure Testing Submission Form

0535_NQF_testing_attachment_Sep2017_3.28.18F.docx

2.1 <u>For maintenance of endorsement</u>

Reliability testing: If testing of reliability of the measure score was not presented in prior submission(s), has reliability testing of the measure score been conducted? If yes, please provide results in the Testing attachment. Please use the most current version of the testing attachment (v7.1). Include information on all testing conducted (prior testing as well as any new testing); use red font to indicate updated testing.

No

2.2 <u>For maintenance of endorsement</u>

Has additional empirical validity testing of the measure score been conducted? If yes, please provide results in the Testing attachment. Please use the most current version of the testing attachment (v7.1). Include information on all testing conducted (prior testing as well as any new testing); use red font to indicate updated testing.

No

2.3 <u>For maintenance of endorsement</u>

Risk adjustment: For outcome, resource use, cost, and some process measures, risk-adjustment that includes social risk factors is not prohibited at present. Please update sections 1.8, 2a2, 2b1,2b4.3 and 2b5 in the Testing attachment and S.140 and S.11 in the online submission form. NOTE: These sections must be updated even if social risk factors are not included in the risk-adjustment strategy. You MUST use the most current version of the Testing Attachment (v7.1) -- older versions of the form will not have all required questions.

Yes - Updated information is included

Measure Testing (subcriteria 2a2, 2b1-2b6)

Measure Number (*if previously endorsed*)**:** 0535
Measure Title: 30-day all-cause risk-standardized mortality rate following Percutaneous Coronary Intervention (PCI) for patients without ST segment elevation myocardial infarction (STEMI) and without cardiogenic shock
Date of Submission: 11/8/2017

Type of Measure:

☒ **Outcome (*including PRO-PM*)**	☐ **Composite – *STOP – use composite testing form***
☐ Intermediate Clinical Outcome	☐ Cost/resource
☐ Process *(including Appropriate Use)*	☐ Efficiency
☐ Structure	

Instructions

- Measures must be tested for all the data sources and levels of analyses that are specified. ***If there is more than one set of data specifications or more than one level of analysis, contact NQF staff*** about how to present all the testing information in one form.
- **For all measures, sections 1, 2a2, 2b1, 2b2, and 2b4 must be completed.**
- **For outcome and resource use measures**, section **2b3** also must be completed.
- If specified for **multiple data sources/sets of specificaitons** (e.g., claims and EHRs), section **2b5** also must be completed.
- Respond to all questions as instructed with answers immediately following the question. All information on testing to demonstrate meeting the subcriteria for reliability (2a2) and validity (2b1-2b6) must be in this form. An appendix for *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 25 pages (*incuding questions/instructions;* minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.
- 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] 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] 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]

AND

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

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; [14,15] 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] **differences in performance**;

OR

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

2b5. If multiple data sources/methods are specified, there is demonstration they produce comparable results.

2b6. Analyses identify the extent and distribution of **missing data** (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias.

Notes

10. Reliability testing applies to both the data elements and computed measure score. Examples of reliability testing for data elements include, but are not limited to: inter-rater/abstractor or intra-rater/abstractor studies; internal consistency for multi-item scales; test-retest for survey items. Reliability testing of the measure score addresses precision of measurement (e.g., signal-to-noise).

11. Validity testing applies to both the data elements and computed measure score. Validity testing of data elements typically analyzes agreement with another authoritative source of the same information. Examples of validity testing of the measure score include, but are not limited to: testing hypotheses that the measures scores indicate quality of care, e.g., measure scores are different for groups known to have differences in quality assessed by another valid quality measure or method; correlation of measure scores with another valid indicator of quality for the specific topic; or relationship to conceptually related measures (e.g., scores on process measures to scores on outcome measures). Face validity of the measure score as a quality indicator may be adequate if accomplished through a systematic and transparent process, by identified experts, and explicitly addresses whether performance scores resulting from the measure as specified can be used to distinguish good from poor quality. The degree of consensus and any areas of disagreement must be provided/discussed.

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

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

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

15. With large enough sample sizes, small differences that are statistically significant may or may not be practically or clinically meaningful. The substantive question may be, for example, whether a statistically significant difference of one percentage point in the percentage of patients who received smoking cessation counseling (e.g., 74 percent v. 75 percent) is clinically meaningful; or whether a statistically significant difference of $25 in cost for an episode of care (e.g., $5,000 v. $5,025) is practically meaningful. Measures with overall less-than-optimal performance may not demonstrate much variability across providers.

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

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

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

Measure Specified to Use Data From: (***must be consistent with data sources entered in S.17***)	**Measure Tested with Data From:**
☐ abstracted from paper record	☐ abstracted from paper record
☒ claims	☒ claims
☒ registry	☒ registry
☐ abstracted from electronic health record	☐ abstracted from electronic health record
☐ eMeasure (HQMF) implemented in EHRs	☐ eMeasure (HQMF) implemented in EHRs
☒ other: Source of vital status (e.g. National Death Index)	☒ other: Source of vital status (e.g. National Death Index)

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

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

Medicare Enrollment Database

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

1.3. What are the dates of the data used in testing?

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

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

Measure Specified to Measure Performance of: (***must be consistent with levels entered in item S.20***)	**Measure Tested at Level of:**
☐ individual clinician	☐ individual clinician
☐ group/practice	☐ group/practice
☒ hospital/facility/agency	☒ hospital/facility/agency
☐ health plan	☐ health plan
☐ other:	☐ other:

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

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

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

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

<u>Measure reliability and validity dataset</u>

The measure reliability and validity dataset linked the CathPCI and Medicare Part A claims data from 2010-2011. It included 255,561 admissions to 1,170 hospitals with 127,781 admissions to 1,167 hospitals in one randomly selected sample and 127,780 admissions to 1,167 hospitals in the remaining sample for patients aged 65 years and older. After excluding hospitals with fewer than 25 cases in each sample, the first sample contained 930 hospitals and the second hospital contained 928 hospitals. The linked dataset was used for:

- Data element reliability testing (Section 2a2)
- Measure score validity testing (Section 2b2)
- Measure exclusions testing (Section 2b3)

<u>Data validity (Section 2b2)</u>

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

<u>Risk adjustment dataset (Section 2b4)</u>

We use admissions with PCI in the merged data from 2006. The development sample consisted of 110,529 admissions at 602 hospitals in the No STEMI/no shock cohort.

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

Social risk factors were not used in this risk model for the following reasons. First, as a detailed clinical registry used for quality assessment and improvement, there are not prospective interviews with patients to obtain patient-reported data. Second, the effect of social risk factors may be at either the patient- or the hospital-level. For example, patients with social risk factors (i.e., low income, lack of education, etc.) may have an increased risk of mortality because these patients may have an individual higher risk (patient-level effect) or because patients with social risk factors are more often admitted to hospitals with higher overall mortality rates (hospital-level effect). It is important to note, however, that even in the presence of a significant patient-level effect and absence of a significant hospital-level effect, the increased risk could be partly or entirely due to the quality of care patients receive in the hospital. For example, biased or differential care provided within a hospital to low-income patients as compared to high-income patients would exert its impact at the level of individual patients, and therefore be a patient-level effect. Third, while it may be true that worse social risk factors might be associated with more severe illness at the time of presentation, we had direct access to detailed clinical variables describing the severity of illness and feel that incorporating such factors (e.g. ejection fraction, PCI status, cardiac arrest, highest risk legion, etc.) is a much more accurate means of stratifying risk. Accordingly, we feel that in this model of 30-day AMI mortality for NSTEMI/No Shock patients, given the rich clinical data available through the NCDR CathPCI Registry and linkage to National Death Index data, that social risk factors, which are not readily available, would not likely contribute much improvement to this particular risk model. Additionally, the results of the disparities data (as shown in 1b.4 of the main submission form) suggests that we do not need to incorporate social risk factors in the model or stratify the measure by SDS.

2a2. RELIABILITY TESTING

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

2a2.1. What level of reliability testing was conducted? (*may be one or both levels*)

☒ **Critical data elements used in the measure** (*e.g., inter-abstractor reliability; data element reliability must address ALL critical data elements*)

☒ **Performance measure score** (e.g., *signal-to-noise analysis*)

2a2.2. For each level checked above, describe the method of reliability testing and what it tests (*describe the steps—do not just name a method; what type of error does it test; what statistical analysis was used*)

Data Element Reliability

See Section 2b2 for validity testing of data elements

Measure Score Reliability

To assess reliability of the measure, we examined the extent to which assessments of a hospital using different but randomly selected subsets of patients in the same time period produced similar measures of hospital performance. That is, we took a "test-retest" approach in which hospital performance is measured once using a random subset of patients, then measured again using a second random subset exclusive of the first, and calculated the agreement of the two resulting performance measures across hospitals.

For test-retest reliability of the measure in Medicare FFS patients aged 65 and older, we combined index admissions from two years (2010 and 2011) into a single dataset, randomly sampled half of patients within each hospital, calculated the measure for each hospital, and repeated the calculation using the second half. Thus, we measured each hospital twice, but each measurement is made using an entirely distinct set of patients. To the extent that the calculated measures of these two subsets agree, we have evidence that the measure is reliable. As a metric of agreement we calculated the intra-class correlation coefficient and assessed the values according to conventional standards.

Specifically, we used a combined 2010-2011 sample that had been linked with Medicare FFS claims data, and randomly split it into two approximately equal subsets of patients. We then calculated the RSMR for each hospital for each sample. The agreement of the two RSMRs was quantified for hospitals in each sample using the intra-class correlation. Using two independent samples provides an honest estimate of the measure's reliability, compared with using two random but potentially overlapping samples, which would exaggerate the agreement. Of note, because our final measure is derived using hierarchical logistic regression, a known property of hierarchical logistic regression models is that smaller volume hospitals contribute less ´signal´. As such a split sample using a single measurement period likely introduces extra noise; potentially underestimating the actual test-retest reliability that would be achieved if the measures were reported using additional years of data. Furthermore, the measure is specified for the entire PCI population, but we tested it only in the subset of Medicare FFS patients for whom information about vital status was available. This reduced the cohort available for testing by approximately 40%.

2a2.3. For each level of testing checked above, what were the statistical results from reliability testing? (e.*g., percent agreement and kappa for the critical data elements; distribution of reliability statistics from a signal-to-noise analysis*)

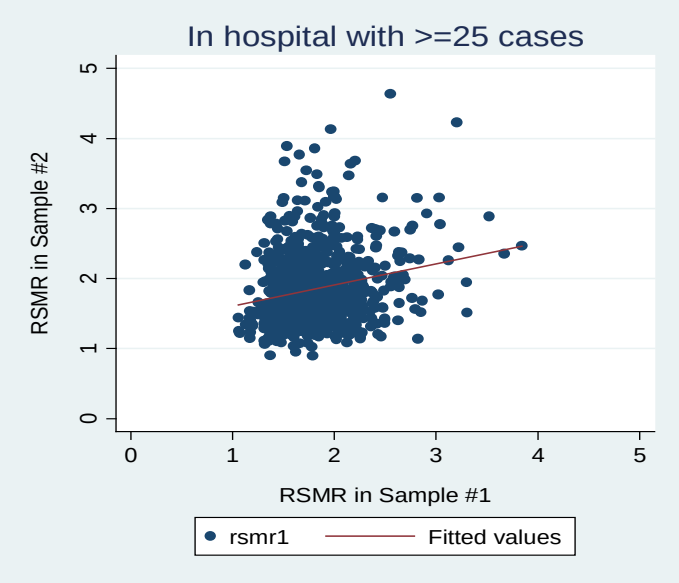
Measure Score Reliability

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

Table 1. Overall mortality rate (OMR) and risk-standardized mortality rate (RSMR) in the split samples; 2010-2011.

Description	First Half of the Data			Second Half of the Data		
	Volume	Weighted by Hospital Volume		Volume	Weighted by Hospital Volume	
		OMR	RSMR		OMR	RSMR
N	1,167	127,781	127,781	1,167	127,780	127,780
Mean	109.50	0.0180	0.0180	109.49	0.0178	0.0179
Std Deviation	116.17	0.0151	0.0051	115.63	0.0143	0.0040
100% Max	1237	0.3333	0.0464	1200	1.0000	0.0384
99%	554	0.0714	0.0338	528	0.0588	0.0312
95%	324	0.0435	0.0273	321	0.0441	0.0255
90%	246	0.0350	0.0242	240	0.0340	0.0229
75% Q3	143	0.0242	0.0207	146	0.0247	0.0200
50% Median	79	0.0155	0.0170	79	0.0159	0.0173
25% Q1	32	0.0088	0.0146	31	0.0080	0.0151
10%	13	0.0000	0.0124	13	0.0000	0.0137
5%	6	0.0000	0.0116	7	0.0000	0.0131
1%	2	0.0000	0.0090	2	0.0000	0.0112
0% Min	1	0.0000	0.0090	1	0.0000	0.0106

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



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

The reliability of a measurement is the degree to which repeated measurements of the same entity agree with each other. For measures of hospital performance, the measured entity is naturally the hospital, and reliability is the extent to

which repeated measurements of the same hospital give similar results. Accordingly, our approach to assessing reliability is to consider the extent to which assessments of a hospital using different but randomly selected subsets of patients in the same time period produce similar measures of hospital performance. The agreement between the two RSMRRs for each hospital was 0.256, which according to the conventional interpretation is "fair" (Landis JR et al. 2013).

References

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

2b1. VALIDITY TESTING

2b1.1. What level of validity testing was conducted? (*may be one or both levels*)

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

☐ **Performance measure score**

☒ **Empirical validity testing**

☐ **Systematic assessment of face validity of performance measure score as an indicator** of quality or resource use (*i.e., is an accurate reflection of performance on quality or resource use and can distinguish good from poor performance*) **NOTE**: Empirical validity testing is expected at time of maintenance review; if not possible, justification is required.

2b1.2. For each level of testing checked above, describe the method of validity testing and what it tests (*describe the steps—do not just name a method; what was tested, e.g., accuracy of data elements compared to authoritative source, relationship to another measure as expected; what statistical analysis was used)*

Data Element Validity

Data element validity testing was done on the specified measure by comparing with variables in the ACC audit program. The NCDR CathPCI Registry has an established Data Quality Program that serves to assess and improve the quality of the data submitted to the registry. There are two complementary components to the Data Quality Program- the Data Quality Report (DQR) and the Data Audit Program (DAP). The DQR process assesses the completeness of the electronic data submitted by participating hospitals. Hospitals must achieve >95% completeness of specific data elements identified as "core fields" to be included in the registry's data warehouse for analysis. The "core fields" encompass the variables included in our risk adjustment models. The process is iterative, providing hospitals with the opportunity to correct errors and resubmit data for review and acceptance into the data warehouse. All data for this analysis passed the DQR completeness thresholds.

The DAP consists of annual on-site chart review and data abstraction. Among participating hospitals that pass the DQR for a minimum of two quarters, at least 5% are randomly selected to participate in the DAP. At individual sites, auditors review charts of 10% of submitted cases. The audits focus on variables that are used in the NCDR risk-adjusted in-hospital mortality model including demographics, comorbidities, cardiac status, coronary anatomy, and PCI status. The DAP includes an appeals process for hospitals to dispute the audit findings. The NCDR DAP was accepted by the National Quality Forum as part of its endorsement of the CathPCI Registry's in-hospital risk-adjusted mortality measure.

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

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

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

Further details also are located in the attached Appendix.

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

Data Element Validity

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

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

Data Element Validity

The audits conducted by the ACC support the overall validity of the data elements included in this measure. The data elements used for risk adjustment were consistently found for all patients and were accurately extracted from the medical record.

2b2. EXCLUSIONS ANALYSIS

NA ☐ no exclusions — *skip to section* 2b4

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

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

2b2.2. What were the statistical results from testing exclusions? (*include overall number and percentage of individuals excluded, frequency distribution of exclusions across measured entities, and impact on performance measure scores*)

Table 2. Exclusions from the target population for the 2010, 2011, and the combined 2010-2011 dataset.

Exclusions	2010		2011		2010-2011
	Patient Stay	Hospitals	Patient Stay	Hospitals	Patient Stay
	# (%)	# (%)	# (%)	# (%)	# (%)
Initial Sample	**199853**	**1,095**	**195,812**	**1,185**	**395,665**
Not Medicare patient on admission	43, 669 (21.85)	3 (0.27)	44,840 (22.90)	1 (0.08)	88,509 (22.37)
Remaining	**156,184**	**1,092**	**150,972**	**1,184**	**307,156**
Not the first claim in the same claim bundle*	3 (0.00)	0 (0.00)	8 (0.01)	0 (0.00)	11 (0.00)
Remaining	**156,181**	**1,092**	**150,964**	**1,184**	**307,145**
Procedure performed more than 10 days after admission	1,074 (0.69)	0 (0.00)	1,212 (0.80)	0 (0.00)	2286 (0.74)
Remaining	**155,107**	**1,092**	**149,752**	**1,184**	**304,859**
Transferred in (PCI to PCI)	186 (0.12)	0 (0.00)	204 (0.14)	(0.00)	390 (0.13)
Remaining	**154,921**	**1,092**	**149,548**	**1,184**	**304,469**
Unknown death	0 (0.00)	0 (0.00%)	0 (0.00)	0 (0.00)	0 (0.00)
Remaining	**154,921**	**1,092**	**149,548**	**1,184**	**304,469**
Duplicate death	65 (0.04)	0 (0.00%)	77 (0.05)	0 (0.00)	142 (0.05)
Remaining	**154,856**	**1,092**	**149,471**	**1,184**	**304,327**
AMA	215 (0.14)	0 (0.00%)	212 (0.14)	0 (0.00)	427 (0.14)
Remaining	**154,641**	**1,092**	**149,259**	**1,184**	**303,900**
With STEMI/Shock	130,942 (84.67)	20 (1.83)	124619 (83.49)	28 (2.36)	255561 (84.09)
Study Sample	**130,942**	**1,068**	**124,619**	**1,154**	**255,561**
Death within 30-days from procedure	23,699 (15.33)		2,315 (1.86)		4,569 (1.79)
In-Hospital death	1,068 (0.82)		1,041 (0.84)		2,109 (0.83)

* Defined as two or more claims in which the admission date of the current claim is before or the same as the discharge date of its previous claim. When this happens, the information at discharge of the first claim is replaced by the information at discharge of the last claim.

2b2.3. What is your interpretation of the results in terms of demonstrating that exclusions are needed to prevent unfair distortion of performance results? (*i.e., the value outweighs the burden of increased data collection and analysis. Note: **If patient preference is an exclusion**, the measure must be specified so that the effect on the performance score is transparent, e.g., scores with and without exclusion*)

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

_____

2b3. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES

If not an intermediate or health outcome, or PRO-PM, or resource use measure, skip to section 2b5.

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

☐ **No risk adjustment or stratification**

☒ **Statistical risk model with 16 risk factors**

☐ **Stratification by risk categories**

☐ **Other,**

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

2b3.2. If an outcome or resource use component measure is not risk adjusted or stratified, provide rationale and analyses to demonstrate that controlling for differences in patient characteristics (case mix) is not needed to achieve fair comparisons across measured entities.

2b3.3a. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors (clinical factors or social risk factors) used in the statistical risk model or for stratification by risk (*e.g., potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of p<0.10; correlation of x or higher; patient factors should be present at the start of care*) **Also discuss any "ordering" of risk factor inclusion**; for example, are social risk factors added after all clinical factors?

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

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

☐ **Published literature**

☐ **Internal data analysis**

☐ **Other (please describe)**

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

To create a model with increased usability while retaining excellent model performance, we tested the performance of the model without those variables considered to be questionably feasible. To select candidate variables, a team of clinicians reviewed all variables in the NCDR CathPCI Registry database (a copy of the data collection form and the complete list of variables collected and submitted by hospitals can be found at www.ncdr.com). We did not consider as candidate variables those that we would not want to adjust for in a quality measure, such as potential complications,

certain patient demographics (e.g., race, socioeconomic status), and patients‟ admission path (e.g., admitted from, or discharged to, a skilled nursing facility [SNF]). Variables were also considered ineligible if they were particularly vulnerable to gaming or were deemed to lack clinical relevance. Based on careful review by a team of clinicians and further informed by a review of the literature, a total of 26 variables were determined to be appropriate for consideration as candidate variables. Our set of candidate variables included two “demographic” variables (age and gender), 15 “history and risk factor” variables, four “cardiac status” variables, one “cath lab visit” variable and four “PCI procedure” variables. The final risk-adjustment model for the NO STEMI/no shock cohort includes 16 variables:

1) Age (10 year increments)

2) Body Mass Index (5 kg/m^2 increments)

3) History of Congestive Heart Failure

4) History of cerebrovascular disease

5) History of peripheral vascular disease

6) History of chronic lung disease

7) Diabetes

None

Non-insulin diabetes

Insulin diabetes

8) Glomerular Filtration Rate (GFR) (derived)

0=Not measured

1="GFR<30"

2="30≤GFR<60"

3=”60≤GFR<90

4="GFR≥90"

9) Previous PCI

10) Heart Failure - current status

11) New York Hospital Association

Class IV

12) Symptom onset

No MI on admission

MI within 24 hours of admission

MI 24+ hours after admission

13) Ejection Fraction percent (EF)

1=Not measured

2="EF<30"

3="30≤ EF<45"

4=“EF≥45”

14) PCI status

1=Elective

2=Urgent

3=Emergency

4=Salvage

15) Highest risk lesion – coronary artery segment category

1=proximal Right Coronary Artery (RCA)/mid Left Anterior Descending (LAD) artery/proximal Circumflex Artery (Cx)

2=proximal LAD

3=Left Main

4= Other

16) Highest risk lesion: Society for Cardiovascular Angiography and Interventions (SCAI)

Class 1

Class 2 or 3

Class 4

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

N/A

2b3.5. Describe the method of testing/analysis used to develop and validate the adequacy of the statistical model or stratification approach (*describe the steps—do not just name a method; what statistical analysis was used*)

Several variables required particular consideration. Variables such as PCI status impart important prognostic information but are vulnerable to systematic misclassification. This is relevant to efforts to publicly report 30-day PCI mortality in that several key variables (e.g., cardiogenic shock and PCI status) may be consistently coded differently across sites. For example, although the CathPCI data dictionary provides detailed definitions of PCI status (http://www.ncdr.com/WebNCDR/ELEMENTS.ASPX), sites may differ in their interpretation of these definitions such that a patient considered an emergent PCI at hospital A may be considered an urgent PCI at hospital B. If differences in coding occur with sufficient frequency, the risk-standardized mortality rate for hospital A might appear lower than hospital B, even if their case mixes and outcomes were otherwise identical.

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

We used logistic regression with stepwise selection (entry $p<0.05$; retention with $p<0.01$) for variable selection. We also assessed the direction and magnitude of the regression coefficients.

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

If stratified, skip to 2b3.9

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

We computed 6 summary statistics for assessing model performance: over-fitting indices, percentage of variation explained by the risk factors, predictive ability, area under the receiver operating characteristic (ROC) curve, distribution of residuals, and model chi-square.

The development model has excellent discrimination, calibration, and fit. The patient-level mortality rate ranges from 0.1% in the lowest predicted decile to 7.0% in the highest predicted decile, a range of 6.9%. The area under the ROC curve is 0.821.

The discrimination and the explained variation of the model at the patient-level are consistent with those of published PCI in-hospital mortality models (Yale-CORE 2008).

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

Table 3. Model Performance: Calibration Results Based on the Logistic Regression Model

Indices	2011 Sample	2010 Sample
Number of Admissions	124,619	130,942
Calibration		
γ0, γ1	0.000, 1.000	0.048, 1.019
ROC	0.807	0.812
Residuals Lack of Fit (Pearson Residual Fall %)		
<-2	0.000	0.000
[-2, 0)	98.142	98.279
[0, 2)	0.092	0.111
[2+	1.765	1.610

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

Table 4. RSMR Model Performance for No STEMI/No Shock Cohort

Indices	Development Sample (2010)	Validation Sample (2011)	Merged Sample (2010-2011)
Number of hospitals	1,068	1,154	1,170
Number of admissions	130,942	124,619	255,561
RSMR			
100% Max	0.0396	0.0471	0.0422
99%	0.0293	0.0337	0.0337
95%	0.0231	0.0284	0.0260
90%	0.0217	0.0260	0.0236
75%	0.0188	0.0217	0.0202
50% Median	0.0166	0.0180	0.0171
25%	0.0149	0.0152	0.0148
10%	0.0132	0.0132	0.0129
5%	0.0124	0.0121	0.0121
1%	0.0111	0.0109	0.0103
0% Min	0.0101	0.0096	0.0098

2b3.9. Results of Risk Stratification Analysis:

N/A

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

The C-statistic of 0.821 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.

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

N/A

_____

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

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

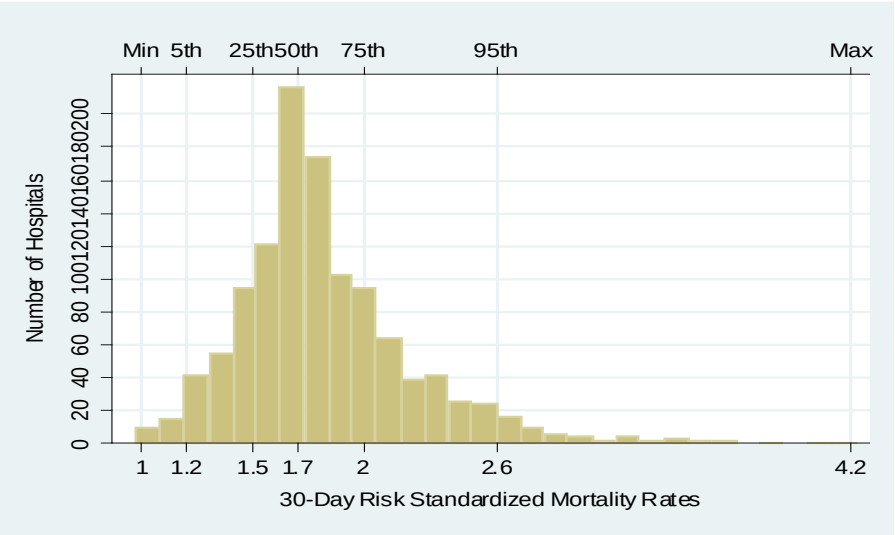
For the currently publicly reported measures of hospital outcomes, including the PCI readmission measure, CMS estimates an interval estimate for each risk-standardized rate to characterize the amount of uncertainty associated with the rate. It then compares the interval estimate to the national crude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

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

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

In the 2010-2011 sample, the mean hospital RSMR for the No STEMI/no shock cohort was 1.8%, with a range of 1.0% to 4.2%. The interquartile range was 1.5% to 2.0%.

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



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

The variation in rates strongly suggests there are meaningful differences across hospitals in the 30-day risk-standardized mortality after PCI in the No STEMI/no shock cohort.

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

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

Note: *This item is directed to measures that are risk-adjusted (with or without social risk factors)* ***OR*** *to measures with more than one set of specifications/instructions (e.g., one set of specifications for how to identify and compute the measure from medical record abstraction and a different set of specifications for claims or eMeasures). It does not apply to measures that use more than one source of data in one set of specifications/instructions (e.g., claims data to identify the denominator and medical record abstraction for the numerator).* ***Comparability is not required when comparing performance scores with and without social risk factors in the risk adjustment model. However, if comparability is not demonstrated for measures with more than one set of specifications/instructions, the different specifications (e.g., for medical records vs. claims) should be submitted as separate measures.***

N/A

2b5.1. Describe the method of testing conducted to compare performance scores for the same entities across the different data sources/specifications (*describe the steps—do not just name a method; what statistical analysis was used*)

N/A

2b5.2. What were the statistical results from testing comparability of performance scores for the same entities when using different data sources/specifications? (*e.g., correlation, rank order*)

N/A

2b5.3. What is your interpretation of the results in terms of the differences in performance measure scores for the same entities across the different data sources/specifications? (i.*e., what do the results mean and what are the norms for the test conducted*)

N/A

2b6. MISSING DATA ANALYSIS AND MINIMIZING BIAS

2b6.1. Describe the method of testing conducted to identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias (*describe the steps—do not just name a method; what statistical analysis was used*)

We examined rates of missing data for all candidate variables and examined histograms of the frequency of missingness by hospital.

2b6.2. What is the overall frequency of missing data, the distribution of missing data across providers, and the results from testing related to missing data? (*e.g., results of sensitivity analysis of the effect of various rules for missing data/nonresponse; if no empirical sensitivity analysis, identify the approaches for handling missing data that were considered and pros and cons of each*)

Overall the percentage of missing values for all categorical variables was very small (<1%). There were three continuous variables with significant numbers of missing values: body mass index (BMI), glomerular filtration rate (GFR), and left ventricular ejection fraction (LVEF). The frequency of missingness by hospital appeared to be evenly distributed across hospitals. Model performance and estimates of hospital RSMR were not significantly different when repeated excluding cases with missing data. The fact that the data was missing did not appear to be at random in that patients with missing data regarding GFR, and LVEF were at higher risk of death than those without missing data. Accordingly we created a dummy variable to capture that information.

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

2b6.3. What is your interpretation of the results in terms of demonstrating that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias**?** (i*.e., what do the results mean in terms of supporting the selected approach for missing data and what are the norms for the test conducted; if no empirical analysis, provide rationale for the selected approach for missing data*)

As noted above, model performance was comparable when we included or excluded cases with missing data.

3. Feasibility

Extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

3a. Byproduct of Care Processes

For clinical measures, the required data elements are routinely generated and used during care delivery (e.g., blood pressure, lab test, diagnosis, medication order).

3a.1. Data Elements Generated as Byproduct of Care Processes.

Generated or collected by and used by healthcare personnel during the provision of care (e.g., blood pressure, lab value, diagnosis, depression score), Abstracted from a record by someone other than person obtaining original information (e.g., chart abstraction for quality measure or registry), Other

If other: The outcome will be determined from an administrative database such as the National Death Index.

3b. Electronic Sources

The required data elements are available in electronic health records or other electronic sources. If the required data are not in electronic health records or existing electronic sources, a credible, near-term path to electronic collection is specified.

3b.1. To what extent are the specified data elements available electronically in defined fields (*i.e., data elements that are needed to compute the performance measure score are in defined, computer-readable fields*) Update this field for **maintenance of endorsement**.

ALL data elements are in defined fields in electronic health records (EHRs)

3b.2. If ALL the data elements needed to compute the performance measure score are not from electronic sources, specify a credible, near-term path to electronic capture, OR provide a rationale for using other than electronic sources. For **maintenance of endorsement**, if this measure is not an eMeasure (eCQM), please describe any efforts to develop an eMeasure (eCQM).

ACC is in the process of developing a common data dictionary mapped to coded terminology standards with the intent of improving interoperability with EHRs and potentially creation of emeasures.

3b.3. If this is an eMeasure, provide a summary of the feasibility assessment in an attached file or make available at a measure-specific URL. Please also complete and attach the NQF Feasibility Score Card.

Attachment:

3c. Data Collection Strategy

Demonstration that the data collection strategy (e.g., source, timing, frequency, sampling, patient confidentiality, costs associated with fees/licensing of proprietary measures) can be implemented (e.g., already in operational use, or testing demonstrates that it is ready to put into operational use). For eMeasures, a feasibility assessment addresses the data elements and measure logic and demonstrates the eMeasure can be implemented or feasibility concerns can be adequately addressed.

3c.1. Required for maintenance of endorsement. Describe difficulties (as a result of testing and/or operational use of the measure) regarding data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, time and cost of data collection, other feasibility/implementation issues.

IF instrument-based, consider implications for both individuals providing data (patients, service recipients, respondents) and those whose performance is being measured.

Implementation of this measure requires matching of patient data to external data source to determine the outcome endpoint (Death 30 days after PCI). This has resulted in several implementation challenges

- Data Availability: ACC was not able to implement the measure using CMS data as we could not identify a mechanism to use CMS Data for purposes of public reporting. Accordingly we have had to modify implementation strategy, rework our models and match NCDR records to CDC National Death Index (NDI) data.
- Patient Confidentiality: CDC NDI requires direct patient identifiers in order to meet the minimum criteria for matching. Roughly 15% of submitting NCDR sites (based on 2017Q2 CathPCI data) do not submit direct patient identifiers to the registry and are therefore ineligible for NDI matching and cannot participate in this measure.
- Data Cost: Use of the CDC NDI data as the source of vital status requires a substantial investment. For example, obtaining vital status on three years of PCI patients resulted in a cost of~$100,000.
- Data Timeliness: CDC NDI is released on a yearly basis, roughly one year after the calendar year of death along with processing time of the matching process and report generation and the most contemporary data available is over 18 months old.

3c.2. Describe any fees, licensing, or other requirements to use any aspect of the measure as specified *(e.g., value/code set, risk model, programming code, algorithm)*.

This measure was developed and designed to be used across other organizations and by other measure implementers. The fee and licensing information include below is specific to NCDR program requirements:

The ACCF's program the National Cardiovascular Data Registry (NCDR) provides evidence based solutions for cardiologists and other medical professionals committed to excellence in cardiovascular care. NCDR hospital participants

receive confidential benchmark reports that include access to measure macro specifications and micro specifications, the eligible patient population, exclusions, and model variables (when applicable). In addition to hospital sites, NCDR Analytic and Reporting Services provides consenting hospitals' aggregated data reports to interested federal and state regulatory agencies, multi-system provider groups, third-party payers, and other organizations that have an identified quality improvement initiative that supports NCDR-participating facilities. Lastly, the ACCF also allows for licensing of the measure specifications outside of the Registry. For calendar year 2014 the annual pricing for hospitals, NCDR Analytic and Reporting Services, and licensing of measure specifications ranges from $2900-$50,000.

Measures that are aggregated by ACCF and submitted to NQF are intended for public reporting and therefore there is no charge for a standard export package. However, on a case by case basis, requests for modifications to the standard export package will be available for a separate charge.

There is no added procedural risk to patients through their hospital's involvement in the CathPCI Registry. No testing, time, risk, or procedures beyond those required for routine care will be imposed.

4. Usability and Use

Extent to which potential audiences (e.g., consumers, purchasers, providers, policy makers) are using or could use performance results for both accountability and performance improvement to achieve the goal of high-quality, efficient healthcare for individuals or populations.

4a. Accountability and Transparency

Performance results are used in at least one accountability application within three years after initial endorsement and are publicly reported within six years after initial endorsement (or the data on performance results are available). If not in use at the time of initial endorsement, then a credible plan for implementation within the specified timeframes is provided.

4.1. Current and Planned Use

NQF-endorsed measures are expected to be used in at least one accountability application within 3 years and publicly reported within 6 years of initial endorsement in addition to performance improvement.

Specific Plan for Use	Current Use (for current use provide URL)
Public Reporting	Quality Improvement (external benchmarking to organizations) National cardiovascular data Registry https://www.ncdr.com/webncdr/cathpci/home/datacollection

4a1.1 For each CURRENT use, checked above (update for maintenance of endorsement), provide:

- Name of program and sponsor
- Purpose
- Geographic area and number and percentage of accountable entities and patients included
- Level of measurement and setting

N/A

4a1.2. If not currently publicly reported OR used in at least one other accountability application (e.g., payment program, certification, licensing) what are the reasons? (*e.g., Do policies or actions of the developer/steward or accountable entities restrict access to performance results or impede implementation?*)

This measure is designed for use in public reporting, but it is currently not in use. See below for rationale and plan for public reporting. ACC plans to include this measure in NCDR's public reporting program in the future.

4a1.3. If not currently publicly reported OR used in at least one other accountability application, provide a credible plan for implementation within the expected timeframes -- any accountability application within 3 years and publicly reported within 6 years of initial endorsement. (*Credible plan includes the specific program, purpose, intended*

audience, and timeline for implementing the measure within the specified timeframes. A plan for accountability applications addresses mechanisms for data aggregation and reporting.)

Update to credible plan (11/8/17):

We moved forward with implementing the 30-day risk adjusted mortality measures in the CathPCI registry for the ´Quarter 3, 2017 30-Day mortality outcomes report´ which included data from 2011 to 2014. However, ACC held off on public reporting since we are also in the process of updating the CathPCI registry to version 5. The new registry version includes elements to assess out-of-hospital cardiac arrest, which has been identified in the literature as a risk factor that should be considered in mortality modeling(1,2). Additionally, when preparing the public reporting metric for in-hospital mortality (#0133) and 30-day mortality (#0536), we found that the measures were not harmonized in structure (i.e. the 30-day measure is a hierarchical model whereas the in-hospital measure is not). As such, these measures could not be rolled up together to create an appropriate composite view of mortality. We plan to modify the in-hospital mortality model to a hierarchical structure when we expand to take advantage of the additional elements in version 5 of CathPCI registry, particularly cardiac arrest, rather than sequencing a number of major revisions in a relatively short time period for hospitals. In order to avoid unintended negative consequences, ACC has made the decision to put a hold on public reporting until the cardiac arrest elements can be considered for modeling and the inpatient and 30-day PCI mortality models can be structurally harmonized. In addition, for purposes of public reporting this measure will also always be paired with (#0536) 30-day all-cause risk-standardized mortality rate following percutaneous coronary intervention (PCI) for patients with ST segment elevation myocardial infarction (STEMI) and with cardiogenic shock.

Additionally, ACC was in the process of transferring measure stewardship from CMS during the last re-endorsement in 2014. Therefore, ACC previously had limited control over the public use of the measure until it resumed primary stewardship of the measure less than 3 years ago

Citation:

[1] Peberdy, M.A., Donnino, M.W., Callaway, C.W., et al. Impact of Percutaneous Coronary Intervention Performance Reporting on Cardiac Resuscitation Centers: A Scientific Statement From the American Heart Association. Circulation. 2013;128:762-773; originally published online July 15, 2013; doi: 10.1161/CIR.0b013e3182a15cd2

[2] Camuglia, A.C., Randhawa, V.K., Lavi, S., et al. Cardiac catheterization is associated with superior outcomes for survivors of out of hospital cardiac arrest: Review and meta-analysis. Elsevier: Resuscitation 85 (2014) 1533–1540 . www.elsevier.com/locate/resuscitation

NCDR Public Reporting Background:

ACC's National Cardiovascular Data Registry (NCDR) Voluntary Hospital Public Reporting Program:

The ACC currently runs a program to give hospitals the opportunity to voluntarily publicly report their measure results based on data from the National Cardiovascular Data Registry (NCDR). Hospitals that choose to participate have their results displayed on ACC's CardioSmart. Currently Hospitals can report on five measures from the CathPCI Registry and five measures from the ICD Registry. Of these publicly reporting measures, five are NQF-endorsed:

- NQF # 1522: Use of a medicine in the ACEi or ARB class to improve heart function after ICD implant in patients with less than normal heart function.
- NQF # 1528: Use of a beta-blocker medication after ICD implant in patients with a previous heart attack.
- NQF #1529: Use of a beta-blocker medication after ICD implant in patients with less than normal heart function.
- NQF #0965: Use of all recommended medications (ACEI or ARB and beta-blocker) to improve heart function and blood pressure after ICD implant.
- NQF # 0964: Therapy with aspirin, P2Y12 inhibitor, and statin at discharge following PCI in eligible patients (composite measure)

4a2.1.1. Describe how performance results, data, and assistance with interpretation have been provided to those being measured or other users during development or implementation.

How many and which types of measured entities and/or others were included? If only a sample of measured entities were included, describe the full population and how the sample was selected.

Performance results are distributed to all CathPCI registry participants as part of quarterly benchmark reports, which provide a detailed analysis of an institution´s individual performance in comparison to the entire registry population from participating hospitals across the nation. Reports include an executive summary dashboard, at-a-glance assessments, and patient level drill-downs. Registry participants also have access to an outcome report companion guide which provides common definitions and detailed metric specifications to assist with interpretation of performance rates.

4a2.1.2. Describe the process(es) involved, including when/how often results were provided, what data were provided, what educational/explanatory efforts were made, etc.

The majority of the required data elements are routinely generated and acquired during the delivery of standard cardiac care to this patient population. Electronic extraction of data recorded as part of the procedure expedites data collection. This strategy offers point of care collection and minimizes time and cost. Institutions can manually report using a free web-based tool or automate the reporting by using certified software developed by third-party vendors. The data elements required for this measure are readily available within the patient's medical record or can be attained without undue burden within the hospital. Most data elements exist in a structured format within patient's electronic health record.

There are a number of methods used to educate and provide general support to registry participants. This includes the following:

- Registry Site Manager Calls are available for all NCDR participants. RSM calls are provided as a source of communication between NCDR and participants to provide a live chat Q and A session on a continuous basis.
- New User Calls are available for NCDR participants, and are intended for assisting new users with their questions.
- NCDR Annual Conference

The NCDR Annual Conference is a well-attended and energetic two-day program at which participants from across the country come together to hear about new NCDR and registry-specific updates. During informative general sessions, attendees can learn about topics such as transcatheter therapies, the NCDR dashboard, risk models, data quality and validation, and value-based purchasing. Attendees also receive registry updates and participate in advanced case studies covering such topics as Appropriate Use Criteria and outcomes report interpretation.

- Release notes (for outcomes reports)
- Clinical Support

The NCDR Product Support and Clinical Quality Consultant Teams are available to assist participating sites with questions Monday through Friday, 9:00 a.m. - 5:00 p.m. ET.

4a2.2.1. Summarize the feedback on measure performance and implementation from the measured entities and others described in 4d.1.

Describe how feedback was obtained.

Feedback is typically obtained through monthly registry site manager monthly calls, ad hoc phone calls tracked with salesforce software, and during registry –specific break-out sessions at the NCDR's annual meeting. Registry Steering Committee members may also provide feedback during regularly scheduled calls.

4a2.2.2. Summarize the feedback obtained from those being measured.

While the 30-day mortality measure was implemented for the first time in the CathPCI registry in Quarter 3, 2017, the registry participants appear to be very interested in this measureThere have been no major issues or other feedback received from registry participants with respect to collecting data for this particular metric.

4a2.2.3. Summarize the feedback obtained from other users

No other feedback was received from other users.

4a2.3. Describe how the feedback described in 4a2.2.1 has been considered when developing or revising the measure specifications or implementation, including whether the measure was modified and why or why not.

N/A (Measure was not modified since last endorsement)

Improvement

Progress toward achieving the goal of high-quality, efficient healthcare for individuals or populations is demonstrated. If not in use for performance improvement at the time of initial endorsement, then a credible rationale describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.

4b1. Refer to data provided in 1b but do not repeat here. Discuss any progress on improvement (trends in performance results, number and percentage of people receiving high-quality healthcare; Geographic area and number and percentage of accountable entities and patients included.)

If no improvement was demonstrated, what are the reasons? If not in use for performance improvement at the time of initial endorsement, provide a credible rationale that describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.

The performance data used and described in 1b reflects a different cohort of data from when the measure was last endorsed. We previously analyzed CMS and CathPCI registry data from 2010 to 2011, however, for this endorsement period had access to the National Death Index (NDI) data from 2011-2014. NDI data is more comprehensive and allowed for the risk model to be applied to all-payers and a wider age range of patients (>18 years of age) compared to CMS data (>65 years of age). Based on the differences in cohorts of data analyzed (CMS vs NDI), we are unable to comment on or draw conclusions from risk adjusted performance trends over time. However, the unadjusted 30-day mortality rate remained low and increased slightly during the study period, from 1.00% in 2011-12 to 1.10% in 2012-2013 and to 1.12% in 2013-14.

4b2. Unintended Consequences

The benefits of the performance measure in facilitating progress toward achieving high-quality, efficient healthcare for individuals or populations outweigh evidence of unintended negative consequences to individuals or populations (if such evidence exists).

4b2.1. Please explain any unexpected findings (positive or negative) during implementation of this measure including unintended impacts on patients.

Ensuring data quality is critical so that the RSMRs can provide fair and accurate estimates of outcomes across hospitals. However, all data sources are potentially prone to misclassifications. Accordingly, adequate mechanisms will need to be implemented to ensure data quality (such as monitoring data for variances in case mix [e.g., unexpectedly high proportion of salvage PCI or cardiogenic shock], chart audits, and possibly adjudicating cases that are vulnerable to systematic misclassification). The NCDR CathPCI registry has successfully implemented methods to ensure that quality data are used for the risk adjustment methodology.

Studies suggest that public reporting of the outcomes of cardiovascular procedures may have unintended consequences. Joynt and colleagues compared the characteristics and outcomes of patients undergoing PCI in states with (MA, NY, PA) and regional states without (CT, DE, ME, MD, NH, RI, VT) public reporting and found that the odds of receiving a PCI for NSTEMI patients in public reporting states versus non-reporting states were similar, whereas STEMI patients with acute MI were less likely to receive PCI in public reporting states than in non-public reporting states. There were no differences in overall 30-day mortality rates among acute MI NSTEMI or STEMI patients in reporting versus non-reporting states. Determining the underlying causes and appropriateness of these differences is impossible, but there is concern that physicians in states that publicly report PCI outcomes would either refer high risk cases to states without public reporting or avoid such cases altogether. Implementing a national measure of PCI outcomes would avoid the former problem in that public reporting would be consistent across states.

Nevertheless, this measure will continue to require close attention to the possibility that high risk patients are not receiving PCI when clinically indicated. The measure is, however, complementary to the previously approved measures for 30-day mortality of AMI and heart failure patients in that inappropriate avoidance of high risk PCI cases may have a detrimental effect on hospitals' performance on these other measures of cardiovascular outcomes. However, it is important to note that this measure has not undergone public reporting to date, thus the unintended consequences are speculative.

Measure implementation will require close attention to data quality. Potential solutions include a) detailed chart audits, b) close attention to variances in case mix and c) review of some or all cases coded as cardiogenic shock or a salvage PCI.

Joynt, K. E., Blumenthal, D. M., Orav, E. J., Resnic, F. S., & Jha, A. K. (2012). Association of Public Reporting for Percutaneous Coronary Intervention with Utilization and Outcomes among Medicare beneficiaries with Acute Myocardial Infarction. JAMA, 308(14), 1460–1468. http://doi.org/10.1001/jama.2012.12922

4b2.2. Please explain any unexpected benefits from implementation of this measure.

N/A - there were no unexpected benefits noted for this measure.

5. Comparison to Related or Competing Measures

If a measure meets the above criteria and there are endorsed or new related measures (either the same measure focus or the same target population) or competing measures (both the same measure focus and the same target population), the measures are compared to address harmonization and/or selection of the best measure.

5. Relation to Other NQF-endorsed Measures

Are there related measures (conceptually, either same measure focus or target population) or competing measures (conceptually both the same measure focus and same target population)? If yes, list the NQF # and title of all related and/or competing measures.

Yes

5.1a. List of related or competing measures (selected from NQF-endorsed measures)

0229 : Hospital 30-day, all-cause, risk-standardized mortality rate (RSMR) following heart failure (HF) hospitalization

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

0536 : 30-day all-cause risk-standardized mortality rate following Percutaneous Coronary Intervention (PCI) for patients with ST segment elevation myocardial infarction (STEMI) or cardiogenic shock

5.1b. If related or competing measures are not NQF endorsed please indicate measure title and steward.

NQF # 0536 - 30-day all-cause risk-standardized mortality rate following percutaneous coronary intervention (PCI) for patients with ST segment elevation myocardial infarction (STEMI) or with cardiogenic shock

NQF # 0230 - Acute Myocardial Infarction 30-day Mortality

NQF # 0229 - Heart Failure 30-day Mortality

5a. Harmonization of Related Measures

The measure specifications are harmonized with related measures;

OR

The differences in specifications are justified

5a.1. If this measure conceptually addresses EITHER the same measure focus OR the same target population as NQF-endorsed measure(s):

Are the measure specifications harmonized to the extent possible?

Yes

5a.2. If the measure specifications are not completely harmonized, identify the differences, rationale, and impact on interpretability and data collection burden.

N/A

5b. Competing Measures

The measure is superior to competing measures (e.g., is a more valid or efficient way to measure);

OR

Multiple measures are justified.

5b.1. If this measure conceptually addresses both the same measure focus and the same target population as NQF-endorsed measure(s):

Describe why this measure is superior to competing measures (e.g., a more valid or efficient way to measure quality); OR provide a rationale for the additive value of endorsing an additional measure. (Provide analyses when possible.)
This measure is most similar to the 30-day all-cause risk-standardized mortality rate following percutaneous coronary intervention (PCI) for patients with ST segment elevation myocardial infarction (STEMI) and with cardiogenic shock. Its additive value stems from the target population of without STEMI and without shock patients.

Appendix

A.1 Supplemental materials may be provided in an appendix. All supplemental materials (such as data collection instrument or methodology reports) should be organized in one file with a table of contents or bookmarks. If material pertains to a specific submission form number, that should be indicated. Requested information should be provided in the submission form and required attachments. There is no guarantee that supplemental materials will be reviewed.

Attachment Attachment: 0535F_Main_Submission_Form_Supplement_04.09.18F.pdf

Contact Information

Co.1 Measure Steward (Intellectual Property Owner): American College of Cardiology

Co.2 Point of Contact: Kim, Lavin, comment@acc.org, 202-375-6448-

Co.3 Measure Developer if different from Measure Steward: American College of Cardiology

Co.4 Point of Contact: Esteban, Perla, eperla@acc.org, 202-375-6499-

Additional Information

Ad.1 Workgroup/Expert Panel involved in measure development

Provide a list of sponsoring organizations and workgroup/panel members' names and organizations. Describe the members' role in measure development.

The measure developer, Yale New Haven Health Servicec Corporation Center for Outcomes Research and Evaluation (YNHHSC/CORE) obtained expert and stakeholder input on the two measures through two mechanisms. First, the team has held regular conference calls with a Working Group of YNHHSC/CORE and American College of Cardiology (ACC)/National Cardiovascular Data Registry (NCDR) experts in cardiovascular registries and in the outcomes measure field. Second, YNHHSC/CORE sought and considered the input of an American College of Cardiology Foundation (ACCF) designated Task Force.

Working Group

Ralph Brindis, M.D., M.P.H., F.A.C.C.

Regional Senior Advisor for Cardiovascular Disease, Northern California Kaiser Permanente; Clinical Professor of Medicine, UCSF, Oakland, CA; Chief Medical Officer and Chairman, Management Board, National Cardiovascular Data Registry

Barbara Christensen, R.N., M.H.A.

Senior Director, Registry Services, American College of Cardiology

Jeptha Curtis, M.D.

Assistant Professor of Medicine, Department of Internal Medicine (Cardiovascular Disease), Yale University

Elizabeth Drye, M.D., S.M.

Research Project Director, Yale/Yale-New Haven Hospital Center for Outcomes Research and Evaluation

Susan Fitzgerald, R.N., M.B.A.

Associate Director, Registry Development, American College of Cardiology

Lori Geary, M.P.H.

Research Project Coordinator, Yale/Yale-New Haven Hospital Center for Outcomes Research and Evaluation

Amy Heller, Ph.D., M.P.H.

Associate Director, Quality Products, American College of Cardiology

Tony Hermann, R.N., M.B.A., C.P.H.Q.

Associate Director, CathPCI Registry, American College of Cardiology

Kathleen Hewitt, R.N., M.S.N., C.P.H.Q.

Associate Vice President, American College of Cardiology

Harlan Krumholz, M.D., M. Sc., F.A.C.C.

Director, Yale Center for Outcomes Research and Evaluation; Representative, NCDR analytic center; Ex-officio to Task Force

Kristi Mitchell, M.P.H.

Senior Director, Research, Development and Quality Products, American College of Cardiology

Eric Peterson, M.D., M.P.H., F.A.C.C.

Professor of Medicine, Duke University; Director, Cardiovascular Outcomes, Duke Clinical Research Institute, Chapel Hill, NC; Member, NCDR Science Oversight Committee/ Representative, NCDR Analytic Center

John Rumsfeld, M.D., Ph.D., F.A.C.C.

Associate Professor of Medicine, University of Colorado; Clinical Coordinator, VA Ischemic

Lara Slattery, M.H.S.

Associate Director, Research, Development, and Quality Products Department – Registries, Products, and Publishing Division, American College of Cardiology

John Spertus, M.D., M.P.H., F.A.C.C.

Director of Cardiovascular Education and Outcomes Research, Mid America Heart Institute, Kansas City, MO; Member, NCDR Science Oversight Committee/Representative, NCDR analytic center; Chair, American College of Cardiology Foundation Task Force on Public Reporting of Hospital-Level Outcomes Measures

Yongfei Wang, M.S.

Senior Research Analyst, Yale/Yale-New Haven Hospital Center for Outcomes Research and Evaluation

William Weintraub, M.D., F.A.C.C.

Chair, CathPCI Registry Steering Committee; Section Chief, Cardiology, Christiana Care Health Services, Inc., Newark DE

Al Woodward, Ph.D., M.B.A.

Director, Research Services, American College of Cardiology

Task Force

Five Task Force members also serve as members of the Working Group, including:

Ralph G. Brindis, M.D., M.P.H., F.A.C.C.

Harlan Krumholz, M.D., M. Sc., F.A.C.C.

Eric Peterson, M.D., M.P.H., F.A.C.C.

John Rumsfeld, M.D., Ph.D., F.A.C.C.

John Spertus, M.D., M.P.H., F.A.C.C.

Other Task Force members are:

John Brush, M.D., F.A.C.C.

Cardiology Consultants LLC, Norfolk, VA; Chair, Quality Strategic Directions Committee

Vincent J. Bufalino, M.D., F.A.C.C.

Midwest Heart Specialists, Naperville, IL; Co-Chair, ACC Advocacy Committee

Gregory Dehmer, M.D., F.A.C.C.

Professor of Medicine, Texas A&M College of Medicine, Temple, TX;

Representative, The Society for Cardiovascular Angiography and Interventions

James Dove, M.D., F.A.C.C.

President, American College of Cardiology

President Emeritus, Prairie Cardiovascular Consultants, Ltd., Springfield, IL; President, ACC/ACCF Board of Trustees

Stephen C. Hammill, M.D., F.H.R.S.

Professor of Medicine, Mayo Clinic College of Medicine, Rochester, MN; Representative, Heart Rhythm Society

Frank E Harrell Jr., PhD

Professor of Biostatistics; Department Chair, Vanderbilt University School of Medicine- Department of Biostatistics, Nashville, TN

Barry K. Lewis, D.O., F.A.C.C.

Consultants in Cardiology, P.C., Farmington Hills, MI; Member, Advocacy Committee

William R. Lewis, M.D., F.A.C.C.

Metro Health Medical Center, Cleveland, OH; ACC Ohio Chapter Governor/ACC Board of Governors

Fred Masoudi, M.D., M.S.P.H., F.A.C.C.

Denver Health Medical Center, Denver, CO; Chair, ACC/AHA Task Force on Performance Measures

Andrea M. Russo, M.D. F.A.C.C.

University of Pennsylvania Health System, Philadelphia, PA; Representative, Heart Rhythm Society

Bonnie H. Weiner, M.D., F.S.C.A.I., F.A.C.C.

Professor of Medicine; Interim Chair Cardiovascular Medicine, St. Vincent Hospital at Worcester Medical Center, Worchester, MA; Representative, The Society for Cardiovascular Angiography and Interventions

Stuart Winston, D.O., F.A.C.C.

Michigan Heart, P. C., Ann Arbor, MI; ACC Michigan Chapter Governor/ACC Board of Governors

Measure Developer/Steward Updates and Ongoing Maintenance

Ad.2 Year the measure was first released: 2009

Ad.3 Month and Year of most recent revision: 12, 2012

Ad.4 What is your frequency for review/update of this measure? With dataset revisions and based on new evidence.

Ad.5 When is the next scheduled review/update for this measure? 04, 2018

Ad.6 Copyright statement: N/A

Ad.7 Disclaimers: N/A

Ad.8 Additional Information/Comments: Please note that the next scheduled review/update for this measure will occur at the same time as the new version release date of the registry in 2018.

NATIONAL QUALITY FORUM

MEASURE WORKSHEET

This document summarizes the evaluation of the measure as it progresses through NQF's Consensus Development Process (CDP). The information submitted by measure developers/stewards is included after the Brief Measure Information, Preliminary Analysis, and Pre-meeting Public and Member Comments sections.

To navigate the links in the worksheet: Click to go to the link. ALT + LEFT ARROW to return

Purple text represents the responses from measure developers.

Red text denotes developer information that has changed since the last measure evaluation review.

Brief Measure Information

NQF #: 2473e

Corresponding Measures:

De.2. Measure Title: Hybrid hospital 30-day, all-cause, risk-standardized mortality rate (RSMR) following acute myocardial infarction (AMI)

Co.1.1. Measure Steward: Centers for Medicare & Medicaid Services (CMS)

De.3. Brief Description of Measure: This measure estimates a hospital-level 30-day, all-cause, risk-standardized mortality rate (RSMR) for patients discharged from the hospital with a principal discharge diagnosis of acute myocardial infarction (AMI). The outcome is all-cause 30-day mortality, defined as death from any cause within 30 days of the index admission date, including in-hospital death, for AMI patients. The target population is Medicare Fee-for-Service beneficiaries who are 65 years or older.

This Hybrid AMI mortality measure was developed de novo. This measure is harmonized with the Centers for Medicare and Medicaid Services' (CMS's) current publicly reported claims-only measure, hospital 30-day, all-cause, risk-standardized mortality rate (RSMR) following acute myocardial infarction (AMI) (NQF #2473). The measure is referred to as a hybrid because it is CMS's intention to calculate the measure using two data sources: Medicare fee-for-service (FFS) administrative claims and clinical electronic health record (EHR) data.

1b.1. Developer Rationale: The goal of this measure is to improve patient outcomes by providing patients, physicians, and hospitals with information about hospital-level, risk-standardized mortality rates following hospitalization for AMI. Measurement of patient outcomes allows for a broad view of quality of care that encompasses more than what can be captured by individual process-of-care measures. Complex and critical aspects of care, such as communication between providers, prevention of and response to complications, patient safety, and coordinated transitions to the outpatient environment, all contribute to patient outcomes but are difficult to measure by individual process measures. The goal of outcomes measurement is to risk-adjust for patients' conditions at the time of hospital admission and then evaluate patient outcomes. This measure was developed to identify institutions whose performance is better or worse than what would be expected based on their patient case mix, and therefore promote hospital quality improvement and better inform consumers about care quality.

Additionally, AMI mortality is a priority area for outcomes measure development as it is a costly and common condition. Hospital mortality is an outcome that is likely attributable to care processes and is an important outcome for patients. Measuring and reporting mortality rates will inform health care providers about opportunities to improve care, strengthen incentives for quality improvement, and ultimately improve the quality of care received by Medicare patients. The measure will also provide patients with information that could guide their choices. Furthermore, the

measure will increase transparency for consumers and has the potential to lower health care costs associated with mortality.

This Hybrid AMI mortality measure incorporates clinical data elements pulled from the EHR in risk adjustment of the mortality model. Some benefits of including the clinical data elements are:

1. Inclusion of patient-level clinical data related to severity of illness is responsive to providers who continue to express preference for using patient-level clinical data, and provides an opportunity to incorporate clinical data into outcome measures.

2. Hospitals will increasingly use EHR data to assess severity of illness and patients' risk of poor outcomes. This provides an opportunity to align the measure with clinical decision support systems that many providers utilize to alert care teams about patients at increased risk of poor outcomes in real time during the inpatient stay.

3. Collecting a simple core set of clinical data elements that perform well as risk-adjustment variables (for illness severity) across conditions can greatly reduce the cost and effort of future measure development, improve harmonization, and create opportunity for longitudinal assessment of patient status and quality of care across settings.

4. These core clinical data elements will provide measure developers with a standard set of reliable data that can be used as a starting place when building risk-adjustment models for quality measures using clinical data.

Numerator Statement: The outcome is all-cause 30-day mortality, defined as death from any cause within 30 days of the index admission date, including in-hospital death, for patients with a principal discharge diagnosis of AMI.

Denominator Statement: The cohort includes inpatient admissions for Medicare FFS patients 65 years and older who were discharged from non-federal, short-term, acute care hospitals with a principal discharge diagnosis of AMI.

Denominator Exclusions: The mortality measure excludes index hospitalizations that meet any of the following exclusion criteria:

1. Discharged alive on the day of admission or the following day, who were not transferred to another acute care facility;

2. With inconsistent or unknown vital status or other unreliable demographic (age and gender) data;

3. Enrolled in the Medicare hospice program any time in the 12 months prior to the index admission, including the first day of the index admission; or

4. Discharged against medical advice (AMA).

After exclusions #1-4 are applied, the measure randomly selects one index admission per patient per year for inclusion in the cohort so that each episode of care is mutually independent with the same probability of the outcome. Additional admissions within that year are excluded. For each patient, the probability of death increases with each subsequent admission and therefore the episodes of care are not mutually independent. For the three-year combined data, when index admissions occur during the transition between measure reporting periods (June and July of each year) and both are randomly selected for inclusion in the measure, the measure includes only the June admission. July admissions are excluded to avoid assigning a single death to two admissions.

Measure Type: Outcome

Data Source: Claims, Electronic Health Data, Electronic Health Records, Other, Registry Data

Level of Analysis: Facility

IF Endorsement Maintenance – Original Endorsement Date: Sep 08, 2014 **Most Recent Endorsement Date:** Sep 08, 2014

Staff Preliminary Analysis: Maintenance of Endorsement

To maintain NQF endorsement endorsed measures are evaluated periodically to ensure that the measures still meets the NQF endorsement criteria ("maintenance"). The emphasis for maintaining endorsement is focused on how effective the measure is for promoting improvements in quality. Endorsed measures should have some experience from the field to inform the evaluation. The emphasis for maintaining endorsement is noted for each criterion.

Criteria 1: Importance to Measure and Report

1a. Evidence

Maintenance measures – less emphasis on evidence unless there is new information or change in evidence since the prior evaluation.

1a. Evidence. The evidence requirements for a health outcome measure include providing empirical data that demonstrate a relationship between the outcome and at least one healthcare structure, process, intervention, or service; if these data not available, data demonstrating wide variation in performance, assuming the data are from a robust number of providers and results are not subject to systematic bias. For measures derived from patient report, evidence also should demonstrate that the target population values the measured outcome, process, or structure and finds it meaningful.

Evidence Summary or **Summary of prior review in 2014**

No new or updated evidence has been submitted since last review. The developer references two citations (Bradley et at., 2012 and Curry et al., 2011) which identify strategies (ex: having cardiologists on site, the presence of physician and nurse champions, and promoting strong communication and coordination across disciplines and departments) used by hospitals that achieve low AMI mortality rates.

Changes to evidence from last review

☒ **The developer attests that there have been no changes in the evidence since the measure was last evaluated.**

☐ **The developer provided updated evidence for this measure:**

Updates: N/A

Question for the Committee:

- *Is there at least one thing that the provider can do to achieve a change in the measure results?*
- *If derived from patient report, does the target population value the measured outcome and finds it meaningful?*

Guidance from the Evidence Algorithm

Outcome measure → Relationship between outcome and at least once healthcare action demonstrated by empirical data (Box 2) → Pass

Preliminary rating for evidence: ☒ **Pass** ☐ **No Pass**

1b. Gap in Care/Opportunity for Improvement and 1b. Disparities

Maintenance measures – increased emphasis on gap and variation

1b. Performance Gap. The performance gap requirements include demonstrating quality problems and opportunity for improvement.

- The developer provided an analysis of the variation in RSMRs among hospitals in the development dataset from 2009 and the CMS publicly reported national rates for the *claims-based measure* of AMI mortality.
- For maintenance measures, NQF requires performance scores on the *measure as specified* (current and over time) at the specified level of analysis.

Disparities

- The developer analyzed 2009 data used for the development of the measure to determine whether disparities exist.
- For maintenance measures, NQF requires disparities data on the measure as specified (current and over time).

Questions for the Committee:

- *Are the performance rates for the claims-based measure applicable to the hybrid measure?*
- *Is there a gap in care that warrants a national performance measure?*
- *Are you aware of evidence that disparities exist in this area of healthcare?*

Preliminary rating for opportunity for improvement: ☐ **High** ☐ **Moderate** ☐ **Low** ☒ **Insufficient**

RATIONALE: The developer did not provide performance and disparities data on the *measure as specified* – this is required for maintenance of endorsement. Instead the developer provided performance data on a similar measure claims-based measure that is publicly reported by CMS.

Committee Pre-evaluation Comments: Criteria 1: Importance to Measure and Report (including 1a, 1b, 1c)

1a. Evidence:

- The evidence for this measure has always been strong.It might have been helpful for the developers to produce evidence to justify that the additional resources to create the hybrid measure would lead to improvements in the measure.
- There is clear evidence that there is a relationship between quality of care and 30-day mortality risk after acute myocardial infarction.
- Lack of new data since last review in 2014 is concerning, since a lot of more current data should be available.
- The measure is deceptive as its title is only partially accurate. This is only mortality in the Medicare population, although the developers say patients older than 18. Second, AMI definition does not separate STEMI/NSTEMI or Type II MI, which adds immense complexity that the measure itself cannot decipher. Lastly, the risk adjustment is, at best, crude, with only 5 elements. One of which is BP which we are uncertain as to which BP is used in the model (lowest, highest, average?) similarly to troponin (lowest, highest, average?)
- This measure comes from empiric evidence and is based upon an expansion of an existing CMS claims-based measure to include EHR data. The citations provided (Bradley et al, 2012 and curry et al 2011) reflect a limited number of hospitals in each paper.
- There are no issues here. The measure focus is clearly important.
- No new evidence since last review.

1b. Gap in Care/Opportunity for Improvement and 1b. Disparities:

- The measure depends on data from 2009. Nothing more recent has been demonstrated.
- The presented data on disparities are from 2009 with hospitals ranked into quartiles by the proportion of patients who are African American. There was no trend across the quartiles. They repeated the same analysis for dual

eligible and found no trend. However, they didn't provide analysis on the measure as specified. They used a similar CMS measure.

- Last evaluated in 2014. Performance gap probably still exists, but any trend in improvement/decline is not documented.
- Performance gap is really difficulty to ascertain. Indeed this is an e-measure, but only captures certain EMR (Corner is not included). This is a major limitation of this measure.
- The performance gap was developed through an analysis of the variations between hospitals in the CMS database.
- Performance data indicates with 2009 registry data with the caveats given elsewhere in this review, there was hospital-level variability in both unadjusted and risk-standardized mortality. Interquartile range was from 10.3% to 11.1%. Whether or not this represents a gap in care or the inadequacy of case mix adjustment cannot be determined from the information given.
- Disparities data not provided on this version of the measure. Based on 2009 data, the IQR was 10.3%-11.1%, range 9.6%-13.1%, mean 10.8%. Data from 2013-2016 on this measure showed IQR 13.2%-14.1%, range 9.7%-18.0%, mean 13.7%. Very narrow IQR, but substantial range.

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability: Specifications **and** Testing
2b. Validity: Testing; Exclusions**;** Risk-Adjustment**;** Meaningful Differences**;** Comparability Missing Data

2c. For composite measures: empirical analysis support composite approach

Reliability

2a1. Specifications requires the measure, as specified, to produce consistent (reliable) and credible (valid) results about the quality of care when implemented. For maintenance measures – no change in emphasis – specifications should be evaluated the same as with new measures.

2a2. Reliability testing demonstrates if 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 enough to distinguish differences in performance across providers. For maintenance measures – less emphasis if no new testing data provided.

Validity

2b2. Validity testing should demonstrate the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality. For maintenance measures – less emphasis if no new testing data provided.

2b2-2b6. Potential threats to validity should be assessed/addressed.

Composite measures only:

2d. Empirical analysis to support composite construction. Empirical analysis should demonstrate that the component measures add value to the composite and that the aggregation and weighting rules are consistent with the quality construct.

eMeasure Technical Advisor(s) review:

Submitted measure is an HQMF compliant eMeasure	The submitted eMeasure specifications follow the industry accepted format for eMeasure (HL7 Health Quality Measures Format (HQMF)). **HQMF specifications** ☒ **Yes** ☐ **No**
Documentation of HQMF or QDM limitations	Submitted eMeasure contains components that cannot be represented due to limitations of HQMF or QDM and the submission explains the work around for these limitations. This is not an electronic clinical quality measure. The HQMF logic component of this measure is intended to extract electronic clinical data elements for measured population and is then linked with administrative claims data.
Value Sets	The submitted eMeasure specifications uses existing value sets when possible and uses new value sets that have been vetted through the VSAC
Measure logic is unambiguous	Submission does not include test results from a simulated data set demonstrating the measure logic can be interpreted precisely and unambiguously;
Feasibility Testing	The submission contains a feasibility assessment that addresses data element feasibility and is based on assessment by 4 EHR vendors: Epic, Cerner, GE, and Meditech..

Complex measure evaluated by Scientific Methods Panel? ☒ **Yes** ☐ **No**

Evaluators: David Cella, Karen Joynt Maddox, Stephen Horner, Bijan Borah and Joseph Kunisch

Evaluation of Reliability and Validity (and composite construction, if applicable):

Review #1, Review #2, Review #3, Review #4, Review #5

Additional Information regarding Scientific Acceptability Evaluation (*if needed*):

Reviewers were not able to come to consensus so both co-chairs also reviewed the measure. Reviewer concerns centered around the validity of the data. For example, the use of registry data to validate the measure and then EMR data to validate the data elements was a concern.

Questions for the Committee regarding reliability:

- *Do you have any concerns that the measure can be consistently implemented (i.e., are measure specifications adequate)?*
- *The Scientific Methods Panel is satisfied with the reliability testing for the measure. Does the Committee think there is a need to discuss and/or vote on reliability?*

Questions for the Committee regarding validity:

- *Do you have any concerns regarding the validity of the measure (e.g., exclusions, risk-adjustment approach, etc.)?*
- *The Scientific Methods Panel is satisfied with the validity analyses for the measure. Does the Committee think there is a need to discuss and/or vote on validity?*

Preliminary rating for reliability: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Preliminary rating for validity: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Review #1: Scientific Acceptability

Measure Number: 2473

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

Scientific Acceptability: Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. Measures must be judged to meet the subcriteria for both reliability and validity to pass this criterion

Instructions for filling out this form:

- Please complete this form for each measure you are evaluating.
- Please pay close attention to the skip logic directions. ***Directives that require you to skip questions are marked in red font.***
- If you are unable to check a box, please highlight or shade the box for your response.
- You must answer the "overall rating" item for both Reliability and Validity. Also, be sure to answer the composite measure question at the end of the form if your measure is a composite.
- For several questions, we have noted which sections of the submission documents you should ***REFERENCE*** and provided ***TIPS*** to help you answer them.
- ***It is critical that you explain your thinking/rationale if you check boxes that require an explanation.*** Please add your explanation directly below the checkbox in a different font color***.*** Also, feel free to add additional explanation, even if you select a checkbox where an explanation is not requested (if you do so, please type this text directly below the appropriate checkbox).
- ***Please refer to the*** Measure Evaluation Criteria and Guidance document ***(pages 18-24) and the 2-page*** Key Points document ***when evaluating your measures***. This evaluation form is an adaptation of Alogorithms 2 and 3, which provide guidance on rating the Reliability and Validity subcriteria.
- ***Remember*** that testing at either the data element level **OR** the measure score level is accepted for some types of measures, but not all (e.g., instrument-based measures, composite measures), and therefore, the embedded rating instructions may not be appropriate for all measures.
- ***Please base your evaluations solely on the submission materials provided by developers.*** NQF strongly discourages the use of outside articles or other resources, even if they are cited in the submission materials. If you require further information or clarification to conduct your evaluation, please communicate with NQF staff (methodspanel@qualityforum.org).

RELIABILITY

1. Are submitted specifications precise, unambiguous, and complete so that they can be consistently implemented?

   REFERENCE: "MIF_xxxx" document

   NOTE*: NQF staff will conduct a separate, more technical, check of eMeasure (eCQM) specifications, value sets, logic, and feasibility, so no need to consider these in your evaluation.*

   TIPS*: Consider the following: Are all the data elements clearly defined? Are all appropriate codes included? Is the logic or calculation algorithm clear? Is it likely this measure can be consistently implemented?*

   ☒Yes (go to Question #2)

   ☐No (please explain below, and go to Question #2) NOTE that even though ***non-precise specifications should result in an overall LOW rating for reliability***, we still want you to look at the testing results.

2. Was empirical reliability testing (at the data element or measure score level) conducted using statistical tests with the measure as specified?

   REFERENCE: "MIF_xxxx" document for specifications, testing attachment questions 1.1-1.4 and section 2a2

***TIPS**: Check the "NO" box below if: only descriptive statistics are provided; only describes process for data management/cleaning/computer programming; testing does not match measure specifications (i.e., data source, level of analysis, included patients, etc.)*

☒Yes (go to Question #3)

☐No, there is reliability testing information, but *not* using statistical tests and/or not for the measure as specified **OR** there is no reliability testing (please explain below, skip Questions #3-8, then go to Question #9)

3. Was reliability testing conducted with computed performance measure scores for each measured entity?

REFERENCE: "Testing attachment_xxx", section 2a2.1 and 2a2.2

***TIPS**: Answer no if: only one overall score for all patients in sample used for testing patient-level data*

☒Yes (go to Question #4)

☐No (skip Questions #4-5 and go to Question #6)

4. Was the method described and appropriate for assessing the proportion of variability due to real differences among measured entities? *NOTE: If multiple methods used, at least one must be appropriate.*

REFERENCE: Testing attachment, section 2a2.2

***TIPS**: Examples of appropriate methods include signal-to-noise analysis (e.g. Adams/RAND tutorial); random split-half correlation; other accepted method with description of how it assesses reliability of the performance score.*

☒Yes (go to Question #5)

☐No (please explain below, then go to question #5 and rate as INSUFFICIENT)

5. **RATING (score level)** - What is the level of certainty or confidence that the performance measure scores are reliable?

REFERENCE: Testing attachment, section 2a2.2

***TIPS**: Consider the following: Is the test sample adequate to generalize for widespread implementation? Do the results demonstrate sufficient reliability so that differences in performance can be identified?*

☐High (go to Question #6)

☒Moderate (go to Question #6)

☐Low (please explain below then go to Question #6)

☐Insufficient (go to Question #6)

6. Was reliability testing conducted with patient-level data elements that are used to construct the performance measure?

REFERENCE: Testing attachment, section 2a2.

***TIPS**: Prior reliability studies of the same data elements may be submitted; if comparing abstraction to "authoritative source/gold standard" go to Question #9)*

☐Yes (go to Question #7)

☒No (if there is score-level testing that you rated something other than INSUFFICIENT in Question #5, skip questions #7-9, then go to Question #10 (OVERALL RELIABILITY); otherwise, skip questions #7-8 and go to Question #9)

7. Was the method described and appropriate for assessing the reliability of ALL critical data elements?

REFERENCE: Testing attachment, section 2a2.2

***TIPS**: For example: inter-abstractor agreement (ICC, Kappa); other accepted method with description of how it assesses reliability of the data elements*

Answer no if: only assessed percent agreement; did not assess separately for all data elements (at least numerator, denominator, exclusions)

☐Yes (go to Question #8)

☐No (if no, please explain below, then go to Question #8 and rate as INSUFFICIENT)

8. **RATING (data element)** – Based on the reliability statistic and scope of testing (number and representativeness of patients and entities), what is the level of certainty or confidence that the data used in the measure are reliable?

   REFERENCE: Testing attachment, section 2a2

   ***TIPS**: Consider the following: Is the test sample adequate to generalize for widespread implementation? Can data elements be collected consistently?*

   ☐Moderate (skip Question #9 and go to Question #10 (OVERALL RELIABILITY); if score-level testing was NOT conducted, rate Question #10 as MODERATE)

   ☐Low (skip Question #9 and go to Question #10 (OVERALL RELIABILITY); if score-level testing was NOT conducted, rate Question #10 (OVERALL RELIABILITY) as LOW)

   ☐Insufficient (go to Question #9)

9. Was **empirical VALIDITY testing** of patient-level data conducted?

   REFERENCE: testing attachment section 2b1.

   NOTE: Skip this question if empirical reliability testing was conducted and you have rated Question #5 and/or #8 as anything other than INSUFFICIENT)

   TIP:** You should answer this question ONLY if score-level or data element reliability testing was NOT conducted or if the methods used were NOT appropriate. For most measures, NQF will accept data element validity testing in lieu of reliability testing—but **check with NQF staff before proceeding, to verify.

   ☐Yes (go to Question #10 and answer using your rating from data element validity testing – Question #23)

   ☐No (please explain below, go to Question #10 (OVERALL RELIABILITY) and rate it as INSUFFICIENT. Then go to Question #11.)

OVERALL RELIABILITY RATING

10. **OVERALL RATING OF RELIABILITY** taking into account precision of specifications (see Question #1) and all testing results:

    ☐High (NOTE: Can be HIGH only if score-level testing has been conducted)

    ☒Moderate (NOTE: Moderate is the highest eligible rating if score-level testing has not been conducted)

    ☐Low (please explain below) [NOTE: Should rate LOW if you believe specifications are NOT precise, unambiguous, and complete]

    ☐Insufficient (please explain below) [NOTE: For most measure types, testing at both the score level and the data element level is not required, but check with NQF staff]

VALIDITY

ASSESSMENT OF THREATS TO VALIDITY

11. Were potential threats to validity that are relevant to the measure empirically assessed ()?

    REFERENCE: Testing attachment, section 2b2-2b6

***TIPS**: Threats to validity that should be assessed include: exclusions; need for risk adjustment; ability to identify statistically significant and meaningful differences; multiple sets of specifications; missing data/nonresponse.*

☒Yes (go to Question #12)

☐No (please explain below and then go to Question #12) [NOTE that ***non-assessment of applicable threats should result in an overall INSUFFICENT rating for validity***]

12. Analysis of potential threats to validity: Any concerns with measure exclusions?

REFERENCE: Testing attachment, section 2b2.

***TIPS**: Consider the following: Are the exclusions consistent with the evidence? Are any patients or patient groups inappropriately excluded from the measure? Are the exclusions/exceptions of sufficient frequency and variation across providers to be needed (and outweigh the data collection burden)? If patient preference (e.g., informed decisionmaking) is a basis for exclusion, does it impact performance and if yes, is the measure specified so that the information about patient preference and the effect on the measure is transparent?*

☐Yes (please explain below then go to Question #13)

☒No (go to Question #13)

☐Not applicable (i.e., there are no exclusions specified for the measure; go to Question #13)

13. Analysis of potential threats to validity: Risk-adjustment (this applies to <u>all</u> outcome, cost, and resource use measures and "NOT APPLICABLE" is not an option for those measures; the risk-adjustment questions (13a-13c, below) also may apply to other types of measures)

REFERENCE: Testing attachment, section 2b3.

13a. Is a conceptual rationale for social risk factors included? ☒Yes ☐No

13b. Are social risk factors included in risk model? ☒Yes ☐No

13c. Any concerns regarding the risk-adjustment approach?

***TIPS**: Consider the following: **If measure is risk adjusted**: If the developer asserts there is **no conceptual basis** for adjusting this measure for social risk factors, do you agree with the rationale? Are the candidate and final variables included in the risk adjustment model adequately described for the measure to be implemented? Are the candidate and final variables included in the risk adjustment model adequately described for the measure to be implemented? Are all of the risk adjustment variables present at the start of care (if not, do you agree with the rationale)? If social risk factors are not included in the risk-adjustment approach, do you agree with the developer's decision? Is an appropriate risk-adjustment strategy included in the measure (e.g., adequate model discrimination and calibration)? Are all statistical model specifications included, including a "clinical model only" if social risk factors are included in the final model? If a measure is NOT risk-adjusted, is a justification for **not risk adjusting** provided (conceptual and/or empirical)? Is there any evidence that contradicts the developer's rationale and analysis for not risk-adjusting?*

☐Yes (please explain below then go to Question #14)

☒No (go to Question #14)

☐Not applicable (e.g., this is a structure or process measure that is not risk-adjusted; go to Question #14)

14. Analysis of potential threats to validity: Any concerns regarding ability to identify meaningful differences in performance or overall poor performance?

REFERENCE: Testing attachment, section 2b4.

☐Yes (please explain below then go to Question #15)

☒No (go to Question #15)

15. Analysis of potential threats to validity: Any concerns regarding comparability of results if multiple data sources or methods are specified?

    REFERENCE: Testing attachment, section 2b5.

    ☐Yes (please explain below then go to Question #16)

    ☒No (go to Question #16)

    ☐Not applicable (go to Question #16)

16. Analysis of potential threats to validity: Any concerns regarding missing data?

    REFERENCE: Testing attachment, section 2b6.

    ☐Yes (please explain below then go to Question #17)

    ☒No (go to Question #17)

ASSESSMENT OF MEASURE TESTING

17. Was empirical validity testing conducted using the measure as specified and with appropriate statistical tests?

    REFERENCE: Testing attachment, section 2b1.

    ***TIPS**: Answer no if: only face validity testing was performed; only refer to clinical evidence; only descriptive statistics; only describe process for data management/cleaning/computer programming; testing does not match measure specifications (i.e. data, eMeasure, level, setting, patients).*

    ☒Yes (go to Question #18)

    ☐No (please explain below, then skip Questions #18-23 and go to Question #24)

18. Was validity testing conducted with computed performance measure scores for each measured entity?

    REFERENCE: Testing attachment, section 2b1.

    ***TIPS**: Answer no if: one overall score for all patients in sample used for testing patient-level data.*

    ☒Yes (go to Question #19)

    ☐No (please explain below, then skip questions #19-20 and go to Question #21)

19. Was the method described and appropriate for assessing conceptually and theoretically sound hypothesized relationships?

    REFERENCE: Testing attachment, section 2b1.

    ***TIPS**: For example: correlation of the performance measure score on this measure and other performance measures; differences in performance scores between groups known to differ on quality; other accepted method with description of how it assesses validity of the performance score*

    ☒Yes (go to Question #20)

    ☐No (please explain below, then go to Question #20 and rate as INSUFFICIENT)

20. **RATING (measure score)** - Based on the measure score results (significance, strength) and scope of testing (number of measured entities and representativeness) and analysis of potential threats, what is the level of certainty or confidence that the performance measure scores are a valid indicator of quality?

    ☐High (go to Question #21)

    ☒Moderate (go to Question #21)

☐Low (please explain below then go to Question #21)

☐Insufficient (go to Question #21)

21. Was validity testing conducted with patient-level data elements?

REFERENCE: Testing attachment, section 2b1.

***TIPS**: Prior validity studies of the same data elements may be submitted*

☒Yes (go to Question #22)

☐No (if there is score-level testing that you rated something other than INSUFFICIENT in Question #20, skip questions #22-25, and go to Question #26 (OVERALL VALIDITY); otherwise, skip questions #22-23 and go to Question #24)

22. Was the method described and appropriate for assessing the accuracy of ALL critical data elements? *NOTE that data element validation from the literature is acceptable.*

***REFERENCE:** Testing attachment, section 2b1.*

***TIPS**: For example: Data validity/accuracy as compared to authoritative source- sensitivity, specificity, PPV, NPV; other accepted method with description of how it assesses validity of the data elements.*

Answer No if: only assessed percent agreement; did not assess separately for all data elements (at least numerator, denominator, exclusions)

☒Yes (go to Question #23)

☐No (please explain below, then go to Question #23 and rate as INSUFFICIENT)

23. **RATING (data element)** - Based on the data element testing results (significance, strength) and scope of testing (number and representativeness of patients and entities) and analysis of potential threats, what is the level of certainty or confidence that the data used in the measure are valid?

☒Moderate (skip Questions #24-25 and go to Question #26)

☐Low (please explain below, skip Questions #24-25 and go to Question #26)

☐Insufficient (go to Question #24 only if no other empirical validation was conducted OR if the measure has not been previously endorsed; otherwise, skip Questions #24-25 and go to Question #26)

24. Was face validity systematically assessed by recognized experts to determine agreement on whether the computed performance measure score from the measure as specified can be used to distinguish good and poor quality?

NOTE: If appropriate empirical testing has been conducted, then evaluation of face validity is not necessary; you should skip this question and Question 25, and answer Question #26 based on your answers to Questions #20 and/or #23]

REFERENCE: Testing attachment, section 2b1.

***TIPS**: Answer no if: focused on data element accuracy/availability/feasibility/other topics; the degree of consensus and any areas of disagreement not provided/discussed.*

☐Yes (go to Question #25)

☐No (please explain below, skip question #25, go to Question #26 (OVERALL VALIDITY) and rate as INSUFFICIENT)

25. **RATING (face validity)** - Do the face validity testing results indicate substantial agreement that the performance measure score from the measure as specified can be used to distinguish quality AND potential threats to validity are not a problem, OR are adequately addressed so results are not biased?

REFERENCE: Testing attachment, section 2b1.

***TIPS**: Face validity is no longer accepted for maintenance measures unless there is justification for why empirical validation is not possible and you agree with that justification.*

☐Yes (if a NEW measure, go to Question #26 (OVERALL VALIDITY) and rate as MODERATE)

☐Yes (if a MAINTENANCE measure, do you agree with the justification for not conducting empirical testing? If no, go to Question #26 (OVERALL VALIDITY) and rate as INSUFFICIENT; otherwise, rate Question #26 as MODERATE)

☐No (please explain below, go to Question #26 (OVERALL VALIDITY) and rate AS LOW)

OVERALL VALIDITY RATING

26. **OVERALL RATING OF VALIDITY** taking into account the results and scope of all testing and analysis of potential threats.

☐High (NOTE: Can be HIGH only if score-level testing has been conducted)

☒Moderate (NOTE: Moderate is the highest eligible rating if score-level testing has NOT been conducted)

☐Low (please explain below) [NOTE: Should rate LOW if you believe that there are threats to validity and/or threats to validity were not assessed]

☐Insufficient (if insufficient, please explain below) [NOTE: For instrument-based measures and some composite measures, testing at both the score level and the data element level is required; if not conducted, should rate as INSUFFICIENT—please check with NQF staff if you have questions.]

I rated "moderate" but there are concerns raised by another panel member regarding validity that reflect his knowledge regarding lower-than expected AMA discharge rates, and lack of "mention regarding disparate data, missing data, outlier data, and erroneous data due to events like a hemolyzed lab specimen." Finally, any evidence that the detected differences reflect differences in care quality?

FOR COMPOSITE MEASURES ONLY: Empirical analyses to support composite construction

27. What is the level of certainty or confidence that the empirical analysis demonstrates that the component measures add value to the composite and that the aggregation and weighting rules are consistent with the quality construct?

REFERENCE: Testing attachment, section 2c

***TIPS**: Consider the following: Do the component measures fit the quality construct? Are the objectives of parsimony and simplicity achieved while supporting the quality construct?*

☐High

☐Moderate

☐Low (please explain below)

☐Insufficient (please explain below)

Review #2: Scientific Acceptability

Measure Number: 2473

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

Scientific Acceptability: Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. Measures must be judged to meet the subcriteria for both reliability and validity to pass this criterion

Instructions for filling out this form:

- Please complete this form for each measure you are evaluating.
- Please pay close attention to the skip logic directions. ***Directives that require you to skip questions are marked in red font.***
- If you are unable to check a box, please highlight or shade the box for your response.
- You must answer the "overall rating" item for both Reliability and Validity. Also, be sure to answer the composite measure question at the end of the form if your measure is a composite.
- For several questions, we have noted which sections of the submission documents you should ***REFERENCE*** and provided ***TIPS*** to help you answer them.
- ***It is critical that you explain your thinking/rationale if you check boxes that require an explanation.*** Please add your explanation directly below the checkbox in a different font color***.*** Also, feel free to add additional explanation, even if you select a checkbox where an explanation is not requested (if you do so, please type this text directly below the appropriate checkbox).
- ***Please refer to the*** Measure Evaluation Criteria and Guidance document ***(pages 18-24) and the 2-page*** Key Points document ***when evaluating your measures***. This evaluation form is an adaptation of Alogorithms 2 and 3, which provide guidance on rating the Reliability and Validity subcriteria.
- ***Remember*** that testing at either the data element level **OR** the measure score level is accepted for some types of measures, but not all (e.g., instrument-based measures, composite measures), and therefore, the embedded rating instructions may not be appropriate for all measures.
- ***Please base your evaluations solely on the submission materials provided by developers.*** NQF strongly discourages the use of outside articles or other resources, even if they are cited in the submission materials. If you require further information or clarification to conduct your evaluation, please communicate with NQF staff (methodspanel@qualityforum.org).

RELIABILITY

28. Are submitted specifications precise, unambiguous, and complete so that they can be consistently implemented?

    REFERENCE: "MIF_xxxx" document

    NOTE*: NQF staff will conduct a separate, more technical, check of eMeasure (eCQM) specifications, value sets, logic, and feasibility, so no need to consider these in your evaluation.*

    TIPS*: Consider the following: Are all the data elements clearly defined? Are all appropriate codes included? Is the logic or calculation algorithm clear? Is it likely this measure can be consistently implemented?*

    ☒Yes (go to Question #2)

    ☐No (please explain below, and go to Question #2) NOTE that even though ***non-precise specifications should result in an overall LOW rating for reliability***, we still want you to look at the testing results.

29. Was empirical reliability testing (at the data element or measure score level) conducted using statistical tests with the measure as specified?

    REFERENCE: "MIF_xxxx" document for specifications, testing attachment questions 1.1-1.4 and section 2a2

***TIPS**: Check the "NO" box below if: only descriptive statistics are provided; only describes process for data management/cleaning/computer programming; testing does not match measure specifications (i.e., data source, level of analysis, included patients, etc.)*

☒Yes (go to Question #3)

☐No, there is reliability testing information, but *not* using statistical tests and/or not for the measure as specified **OR** there is no reliability testing (please explain below, skip Questions #3-8, then go to Question #9)

30. Was reliability testing conducted with computed performance measure scores for each measured entity?

REFERENCE: "Testing attachment_xxx", section 2a2.1 and 2a2.2

***TIPS**: Answer no if: only one overall score for all patients in sample used for testing patient-level data*

☒Yes (go to Question #4)

☐No (skip Questions #4-5 and go to Question #6)

31. Was the method described and appropriate for assessing the proportion of variability due to real differences among measured entities? *NOTE: If multiple methods used, at least one must be appropriate.*

REFERENCE: Testing attachment, section 2a2.2

***TIPS**: Examples of appropriate methods include signal-to-noise analysis (e.g. Adams/RAND tutorial); random split-half correlation; other accepted method with description of how it assesses reliability of the performance score.*

☒Yes (go to Question #5)

☐No (please explain below, then go to question #5 and rate as INSUFFICIENT)

32. **RATING (score level)** - What is the level of certainty or confidence that the performance measure scores are reliable?

REFERENCE: Testing attachment, section 2a2.2

***TIPS**: Consider the following: Is the test sample adequate to generalize for widespread implementation? Do the results demonstrate sufficient reliability so that differences in performance can be identified?*

☐High (go to Question #6)

☒Moderate (go to Question #6)

☐Low (please explain below then go to Question #6)

☐Insufficient (go to Question #6)

33. Was reliability testing conducted with patient-level data elements that are used to construct the performance measure?

REFERENCE: Testing attachment, section 2a2.

***TIPS**: Prior reliability studies of the same data elements may be submitted; if comparing abstraction to "authoritative source/gold standard" go to Question #9)*

☐Yes (go to Question #7)

☒No (if there is score-level testing that you rated something other than INSUFFICIENT in Question #5, skip questions #7-9, then go to Question #10 (OVERALL RELIABILITY); otherwise, skip questions #7-8 and go to Question #9)

34. Was the method described and appropriate for assessing the reliability of ALL critical data elements?

REFERENCE: Testing attachment, section 2a2.2

***TIPS**: For example: inter-abstractor agreement (ICC, Kappa); other accepted method with description of how it assesses reliability of the data elements*

Answer no if: only assessed percent agreement; did not assess separately for all data elements (at least numerator, denominator, exclusions)

☐Yes (go to Question #8)

☐No (if no, please explain below, then go to Question #8 and rate as INSUFFICIENT)

35. **RATING (data element)** – Based on the reliability statistic and scope of testing (number and representativeness of patients and entities), what is the level of certainty or confidence that the data used in the measure are reliable?

REFERENCE: Testing attachment, section 2a2

***TIPS**: Consider the following: Is the test sample adequate to generalize for widespread implementation? Can data elements be collected consistently?*

☐Moderate (skip Question #9 and go to Question #10 (OVERALL RELIABILITY); if score-level testing was NOT conducted, rate Question #10 as MODERATE)

☐Low (skip Question #9 and go to Question #10 (OVERALL RELIABILITY); if score-level testing was NOT conducted, rate Question #10 (OVERALL RELIABILITY) as LOW)

☐Insufficient (go to Question #9)

36. Was **empirical VALIDITY testing** of patient-level data conducted?

REFERENCE: testing attachment section 2b1.

NOTE: Skip this question if empirical reliability testing was conducted and you have rated Question #5 and/or #8 as anything other than INSUFFICIENT)

TIP:** You should answer this question ONLY if score-level or data element reliability testing was NOT conducted or if the methods used were NOT appropriate. For most measures, NQF will accept data element validity testing in lieu of reliability testing—but **check with NQF staff before proceeding, to verify.

☐Yes (go to Question #10 and answer using your rating from data element validity testing – Question #23)

☐No (please explain below, go to Question #10 (OVERALL RELIABILITY) and rate it as INSUFFICIENT. Then go to Question #11.)

OVERALL RELIABILITY RATING

37. **OVERALL RATING OF RELIABILITY** taking into account precision of specifications (see Question #1) and all testing results:

☐High (NOTE: Can be HIGH only if score-level testing has been conducted)

☒Moderate (NOTE: Moderate is the highest eligible rating if score-level testing has not been conducted)

☐Low (please explain below) [NOTE: Should rate LOW if you believe specifications are NOT precise, unambiguous, and complete]

☐Insufficient (please explain below) [NOTE: For most measure types, testing at both the score level and the data element level is not required, but check with NQF staff]

VALIDITY

ASSESSMENT OF THREATS TO VALIDITY

38. Were potential threats to validity that are relevant to the measure empirically assessed ()?

REFERENCE: Testing attachment, section 2b2-2b6

***TIPS**: Threats to validity that should be assessed include: exclusions; need for risk adjustment; ability to identify statistically significant and meaningful differences; multiple sets of specifications; missing data/nonresponse.*

☒Yes (go to Question #12)

☐No (please explain below and then go to Question #12) [NOTE that ***non-assessment of applicable threats should result in an overall INSUFFICENT rating for validity***]

39. Analysis of potential threats to validity: Any concerns with measure exclusions?

REFERENCE: Testing attachment, section 2b2.

***TIPS**: Consider the following: Are the exclusions consistent with the evidence? Are any patients or patient groups inappropriately excluded from the measure? Are the exclusions/exceptions of sufficient frequency and variation across providers to be needed (and outweigh the data collection burden)? If patient preference (e.g., informed decisionmaking) is a basis for exclusion, does it impact performance and if yes, is the measure specified so that the information about patient preference and the effect on the measure is transparent?*

☐Yes (please explain below then go to Question #13)

☒No (go to Question #13)

☐Not applicable (i.e., there are no exclusions specified for the measure; go to Question #13)

40. Analysis of potential threats to validity: Risk-adjustment (this applies to <u>all</u> outcome, cost, and resource use measures and "NOT APPLICABLE" is not an option for those measures; the risk-adjustment questions (13a-13c, below) also may apply to other types of measures)

REFERENCE: Testing attachment, section 2b3.

13a. Is a conceptual rationale for social risk factors included? ☒Yes ☐No

13b. Are social risk factors included in risk model? ☐Yes ☒No

13c. Any concerns regarding the risk-adjustment approach?

***TIPS**: Consider the following: **If measure is risk adjusted**: If the developer asserts there is **no conceptual basis** for adjusting this measure for social risk factors, do you agree with the rationale? Are the candidate and final variables included in the risk adjustment model adequately described for the measure to be implemented? Are the candidate and final variables included in the risk adjustment model adequately described for the measure to be implemented? Are all of the risk adjustment variables present at the start of care (if not, do you agree with the rationale)? If social risk factors are not included in the risk-adjustment approach, do you agree with the developer's decision? Is an appropriate risk-adjustment strategy included in the measure (e.g., adequate model discrimination and calibration)? Are all statistical model specifications included, including a "clinical model only" if social risk factors are included in the final model? If a measure is NOT risk-adjusted, is a justification for **not risk adjusting** provided (conceptual and/or empirical)? Is there any evidence that contradicts the developer's rationale and analysis for not risk-adjusting?*

☒Yes (please explain below then go to Question #14)

☐No (go to Question #14)

☐Not applicable (e.g., this is a structure or process measure that is not risk-adjusted; go to Question #14)

This could be better-explained. There are only five risk adjusters listed, which do not seem adequate to capture differences in comorbidities that would impact mortality rates. That said, the c-statistic is quite good, so it is possible that they have just elected not to show the clinical comorbidities obtained from claims data? Or the variables they've chosen are powerful ones.

The testing for social risk is done in a different dataset altogether, and despite finding a significant relationship with dual status, the developers feel this should not be included because excluding it does not meaningfully change performance. I assume the measure in which social risk was tested is one with all of the claims-based comorbidities but NOT the physiologic and lab parameters. Therefore, the measure being submitted has not really been evaluated in terms of social risk.

41. Analysis of potential threats to validity: Any concerns regarding ability to identify meaningful differences in performance or overall poor performance?

    REFERENCE: Testing attachment, section 2b4.

    ☐Yes (please explain below then go to Question #15)

    ☒No (go to Question #15)

42. Analysis of potential threats to validity: Any concerns regarding comparability of results if multiple data sources or methods are specified?

    REFERENCE: Testing attachment, section 2b5.

    ☒Yes (please explain below then go to Question #16)

    ☐No (go to Question #16)

    ☐Not applicable (go to Question #16)

I'm confused by the use of registry data to validate the measure and then EMR data to validate the data elements. This may be all that is currently feasible, but it seems important to use the actual data sources from which the measure would be calculated to assess validity. Also very odd that they are using 9-year-old data when much more updated data (registry, EMR, and claims) certainly exist.

43. Analysis of potential threats to validity: Any concerns regarding missing data?

    REFERENCE: Testing attachment, section 2b6.

    ☒Yes (please explain below then go to Question #17)

    ☐No (go to Question #17)

High proportion of missing troponin levels, which may be salient given the small number of risk adjusters included in the clinical data. That said, troponin is such an essential component of AMI care that this seems easily remediable in practice since it is a structured data field in EMRs and often used.

ASSESSMENT OF MEASURE TESTING

44. Was empirical validity testing conducted using the measure as specified and with appropriate statistical tests?

    REFERENCE: Testing attachment, section 2b1.

    ***TIPS**: Answer no if: only face validity testing was performed; only refer to clinical evidence; only descriptive statistics; only describe process for data management/cleaning/computer programming; testing does not match measure specifications (i.e. data, eMeasure, level, setting, patients).*

    ☒Yes (go to Question #18)

    ☐No (please explain below, then skip Questions #18-23 and go to Question #24)

45. Was validity testing conducted with computed performance measure scores for each measured entity?

    REFERENCE: Testing attachment, section 2b1.

    ***TIPS**: Answer no if: one overall score for all patients in sample used for testing patient-level data.*

    ☒Yes (go to Question #19)

☐No (please explain below, then skip questions #19-20 and go to Question #21)

46. Was the method described and appropriate for assessing conceptually and theoretically sound hypothesized relationships?

    REFERENCE: Testing attachment, section 2b1.

    TIPS*: For example: correlation of the performance measure score on this measure and other performance measures; differences in performance scores between groups known to differ on quality; other accepted method with description of how it assesses validity of the performance score*

    ☒Yes (go to Question #20)

    ☐No (please explain below, then go to Question #20 and rate as INSUFFICIENT)

47. **RATING (measure score)** - Based on the measure score results (significance, strength) and scope of testing (number of measured entities and representativeness) and analysis of potential threats, what is the level of certainty or confidence that the performance measure scores are a valid indicator of quality?

    ☐High (go to Question #21)

    ☒Moderate (go to Question #21)

    ☐Low (please explain below then go to Question #21)

    ☐Insufficient (go to Question #21)

48. Was validity testing conducted with patient-level data elements?

    REFERENCE: Testing attachment, section 2b1.

    TIPS*: Prior validity studies of the same data elements may be submitted*

    ☒Yes (go to Question #22)

    ☐No (if there is score-level testing that you rated something other than INSUFFICIENT in Question #20, skip questions #22-25, and go to Question #26 (OVERALL VALIDITY); otherwise, skip questions #22-23 and go to Question #24)

49. Was the method described and appropriate for assessing the accuracy of ALL critical data elements? *NOTE that data element validation from the literature is acceptable.*

    REFERENCE: *Testing attachment, section 2b1.*

    TIPS*: For example: Data validity/accuracy as compared to authoritative source- sensitivity, specificity, PPV, NPV; other accepted method with description of how it assesses validity of the data elements.*

    Answer No if: only assessed percent agreement; did not assess separately for all data elements (at least numerator, denominator, exclusions)

    ☒Yes (go to Question #23)

    ☐No (please explain below, then go to Question #23 and rate as INSUFFICIENT)

50. **RATING (data element)** - Based on the data element testing results (significance, strength) and scope of testing (number and representativeness of patients and entities) and analysis of potential threats, what is the level of certainty or confidence that the data used in the measure are valid?

    ☒Moderate (skip Questions #24-25 and go to Question #26)

    ☐Low (please explain below, skip Questions #24-25 and go to Question #26)

☐Insufficient (go to Question #24 only if no other empirical validation was conducted OR if the measure has not been previously endorsed; otherwise, skip Questions #24-25 and go to Question #26)

51. Was face validity systematically assessed by recognized experts to determine agreement on whether the computed performance measure score from the measure as specified can be used to distinguish good and poor quality?

NOTE: If appropriate empirical testing has been conducted, then evaluation of face validity is not necessary; you should skip this question and Question 25, and answer Question #26 based on your answers to Questions #20 and/or #23]

REFERENCE: Testing attachment, section 2b1.

***TIPS**: Answer no if: focused on data element accuracy/availability/feasibility/other topics; the degree of consensus and any areas of disagreement not provided/discussed.*

☐Yes (go to Question #25)

☐No (please explain below, skip question #25, go to Question #26 (OVERALL VALIDITY) and rate as INSUFFICIENT)

52. **RATING (face validity)** - Do the face validity testing results indicate substantial agreement that the performance measure score from the measure as specified can be used to distinguish quality AND potential threats to validity are not a problem, OR are adequately addressed so results are not biased?

REFERENCE: Testing attachment, section 2b1.

***TIPS**: Face validity is no longer accepted for maintenance measures unless there is justification for why empirical validation is not possible and you agree with that justification.*

☐Yes (if a NEW measure, go to Question #26 (OVERALL VALIDITY) and rate as MODERATE)

☐Yes (if a MAINTENANCE measure, do you agree with the justification for not conducting empirical testing? If no, go to Question #26 (OVERALL VALIDITY) and rate as INSUFFICIENT; otherwise, rate Question #26 as MODERATE)

☐No (please explain below, go to Question #26 (OVERALL VALIDITY) and rate AS LOW)

OVERALL VALIDITY RATING

53. **OVERALL RATING OF VALIDITY** taking into account the results and scope of all testing and analysis of potential threats.

☐High (NOTE: Can be HIGH only if score-level testing has been conducted)

☒Moderate (NOTE: Moderate is the highest eligible rating if score-level testing has NOT been conducted)

☐Low (please explain below) [NOTE: Should rate LOW if you believe that there are threats to validity and/or threats to validity were not assessed]

☐Insufficient (if insufficient, please explain below) [NOTE: For instrument-based measures and some composite measures, testing at both the score level and the data element level is required; if not conducted, should rate as INSUFFICIENT—please check with NQF staff if you have questions.]

FOR COMPOSITE MEASURES ONLY: Empirical analyses to support composite construction

54. What is the level of certainty or confidence that the empirical analysis demonstrates that the component measures add value to the composite and that the aggregation and weighting rules are consistent with the quality construct?

REFERENCE: Testing attachment, section 2c

***TIPS**: Consider the following: Do the component measures fit the quality construct? Are the objectives of parsimony and simplicity achieved while supporting the quality construct?*

☐High

☐Moderate

☐Low (please explain below)

☐Insufficient (please explain below)

Review #3: Scientific Acceptability

Measure Number: 2473

Measure Title: Hybrid Hospital 30-Day All-Cause Risk-Standardized Mortality Rate Following AMI

Scientific Acceptability: Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. Measures must be judged to meet the subcriteria for both reliability and validity to pass this criterion

Instructions for filling out this form:

- Please complete this form for each measure you are evaluating.
- Please pay close attention to the skip logic directions. ***Directives that require you to skip questions are marked in red font.***
- If you are unable to check a box, please highlight or shade the box for your response.
- You must answer the "overall rating" item for both Reliability and Validity. Also, be sure to answer the composite measure question at the end of the form if your measure is a composite.
- For several questions, we have noted which sections of the submission documents you should ***REFERENCE*** and provided ***TIPS*** to help you answer them.
- ***It is critical that you explain your thinking/rationale if you check boxes that require an explanation.*** Please add your explanation directly below the checkbox in a different font color***.*** Also, feel free to add additional explanation, even if you select a checkbox where an explanation is not requested (if you do so, please type this text directly below the appropriate checkbox).
- ***Please refer to the*** Measure Evaluation Criteria and Guidance document ***(pages 18-24) and the 2-page*** Key Points document ***when evaluating your measures***. This evaluation form is an adaptation of Alogorithms 2 and 3, which provide guidance on rating the Reliability and Validity subcriteria.
- ***Remember*** that testing at either the data element level **OR** the measure score level is accepted for some types of measures, but not all (e.g., instrument-based measures, composite measures), and therefore, the embedded rating instructions may not be appropriate for all measures.
- ***Please base your evaluations solely on the submission materials provided by developers.*** NQF strongly discourages the use of outside articles or other resources, even if they are cited in the submission materials. If you require further information or clarification to conduct your evaluation, please communicate with NQF staff (methodspanel@qualityforum.org).

RELIABILITY

55. Are submitted specifications precise, unambiguous, and complete so that they can be consistently implemented?

    REFERENCE: "MIF_xxxx" document

    NOTE*: NQF staff will conduct a separate, more technical, check of eMeasure (eCQM) specifications, value sets, logic, and feasibility, so no need to consider these in your evaluation.*

    TIPS*: Consider the following: Are all the data elements clearly defined? Are all appropriate codes included? Is the logic or calculation algorithm clear? Is it likely this measure can be consistently implemented?*

    ☒Yes (go to Question #2)

    ☐No (please explain below, and go to Question #2) NOTE that even though ***non-precise specifications should result in an overall LOW rating for reliability***, we still want you to look at the testing results.

56. Was empirical reliability testing (at the data element or measure score level) conducted using statistical tests with the measure as specified?

    REFERENCE: "MIF_xxxx" document for specifications, testing attachment questions 1.1-1.4 and section 2a2

***TIPS**: Check the "NO" box below if: only descriptive statistics are provided; only describes process for data management/cleaning/computer programming; testing does not match measure specifications (i.e., data source, level of analysis, included patients, etc.)*

☒Yes (go to Question #3)

☐No, there is reliability testing information, but *not* using statistical tests and/or not for the measure as specified **OR** there is no reliability testing (please explain below, skip Questions #3-8, then go to Question #9)

57. Was reliability testing conducted with computed performance measure scores for each measured entity?

REFERENCE: "Testing attachment_xxx", section 2a2.1 and 2a2.2

***TIPS**: Answer no if: only one overall score for all patients in sample used for testing patient-level data*

☒Yes (go to Question #4)

☐No (skip Questions #4-5 and go to Question #6)

58. Was the method described and appropriate for assessing the proportion of variability due to real differences among measured entities? *NOTE: If multiple methods used, at least one must be appropriate.*

REFERENCE: Testing attachment, section 2a2.2

***TIPS**: Examples of appropriate methods include signal-to-noise analysis (e.g. Adams/RAND tutorial); random split-half correlation; other accepted method with description of how it assesses reliability of the performance score.*

☒Yes (go to Question #5)

☐No (please explain below, then go to question #5 and rate as INSUFFICIENT)

59. **RATING (score level)** - What is the level of certainty or confidence that the performance measure scores are reliable?

REFERENCE: Testing attachment, section 2a2.2

***TIPS**: Consider the following: Is the test sample adequate to generalize for widespread implementation? Do the results demonstrate sufficient reliability so that differences in performance can be identified?*

☐High (go to Question #6)

☒Moderate (go to Question #6)

☐Low (please explain below then go to Question #6)

☐Insufficient (go to Question #6)

60. Was reliability testing conducted with patient-level data elements that are used to construct the performance measure?

REFERENCE: Testing attachment, section 2a2.

***TIPS**: Prior reliability studies of the same data elements may be submitted; if comparing abstraction to "authoritative source/gold standard" go to Question #9)*

☐Yes (go to Question #7)

☒No (if there is score-level testing that you rated something other than INSUFFICIENT in Question #5, skip questions #7-9, then go to Question #10 (OVERALL RELIABILITY); otherwise, skip questions #7-8 and go to Question #9)

61. Was the method described and appropriate for assessing the reliability of ALL critical data elements?

REFERENCE: Testing attachment, section 2a2.2

***TIPS**: For example: inter-abstractor agreement (ICC, Kappa); other accepted method with description of how it assesses reliability of the data elements*

Answer no if: only assessed percent agreement; did not assess separately for all data elements (at least numerator, denominator, exclusions)

☐Yes (go to Question #8)

☐No (if no, please explain below, then go to Question #8 and rate as INSUFFICIENT)

62. **RATING (data element)** – Based on the reliability statistic and scope of testing (number and representativeness of patients and entities), what is the level of certainty or confidence that the data used in the measure are reliable?

REFERENCE: Testing attachment, section 2a2

***TIPS**: Consider the following: Is the test sample adequate to generalize for widespread implementation? Can data elements be collected consistently?*

☐Moderate (skip Question #9 and go to Question #10 (OVERALL RELIABILITY); if score-level testing was NOT conducted, rate Question #10 as MODERATE)

☐Low (skip Question #9 and go to Question #10 (OVERALL RELIABILITY); if score-level testing was NOT conducted, rate Question #10 (OVERALL RELIABILITY) as LOW)

☐Insufficient (go to Question #9)

63. Was **empirical VALIDITY testing** of patient-level data conducted?

REFERENCE: testing attachment section 2b1.

NOTE: Skip this question if empirical reliability testing was conducted and you have rated Question #5 and/or #8 as anything other than INSUFFICIENT)

TIP:** You should answer this question ONLY if score-level or data element reliability testing was NOT conducted or if the methods used were NOT appropriate. For most measures, NQF will accept data element validity testing in lieu of reliability testing—but **check with NQF staff before proceeding, to verify.

☐Yes (go to Question #10 and answer using your rating from data element validity testing – Question #23)

☐No (please explain below, go to Question #10 (OVERALL RELIABILITY) and rate it as INSUFFICIENT. Then go to Question #11.)

OVERALL RELIABILITY RATING

64. **OVERALL RATING OF RELIABILITY** taking into account precision of specifications (see Question #1) and all testing results:

☐High (NOTE: Can be HIGH only if score-level testing has been conducted)

☒Moderate (NOTE: Moderate is the highest eligible rating if score-level testing has not been conducted)

☐Low (please explain below) [NOTE: Should rate LOW if you believe specifications are NOT precise, unambiguous, and complete]

☐Insufficient (please explain below) [NOTE: For most measure types, testing at both the score level and the data element level is not required, but check with NQF staff]

VALIDITY

ASSESSMENT OF THREATS TO VALIDITY

65. Were potential threats to validity that are relevant to the measure empirically assessed ()?

REFERENCE: Testing attachment, section 2b2-2b6

***TIPS**: Threats to validity that should be assessed include: exclusions; need for risk adjustment; ability to identify statistically significant and meaningful differences; multiple sets of specifications; missing data/nonresponse.*

☒Yes (go to Question #12)

☐No (please explain below and then go to Question #12) [NOTE that ***non-assessment of applicable threats should result in an overall INSUFFICENT rating for validity***]

66. Analysis of potential threats to validity: Any concerns with measure exclusions?

REFERENCE: Testing attachment, section 2b2.

***TIPS**: Consider the following: Are the exclusions consistent with the evidence? Are any patients or patient groups inappropriately excluded from the measure? Are the exclusions/exceptions of sufficient frequency and variation across providers to be needed (and outweigh the data collection burden)? If patient preference (e.g., informed decisionmaking) is a basis for exclusion, does it impact performance and if yes, is the measure specified so that the information about patient preference and the effect on the measure is transparent?*

☒Yes (please explain below then go to Question #13)

☐No (go to Question #13)

☐Not applicable (i.e., there are no exclusions specified for the measure; go to Question #13)

- The measure author indicates patients discharged alive on the day of admission or the following day but not transferred to another acute care facility are excluded but no rationale for the exclusion is provided.
- The measure author indicates that patients with one than more admission for a given condition in a given year, only one index admission for that condition is randomly selected for inclusion in the cohort but no rationale for this approach is given.
- The measure author does not exclude patients with secondary conditions that are likely to result in readmissions that are unrelated to the AMI such as cancer or behavioral health conditions.
- The measure does not include a provision for excluding planned readmissions.

67. Analysis of potential threats to validity: Risk-adjustment (this applies to <u>all</u> outcome, cost, and resource use measures and "NOT APPLICABLE" is not an option for those measures; the risk-adjustment questions (13a-13c, below) also may apply to other types of measures)

REFERENCE: Testing attachment, section 2b3.

13a. Is a conceptual rationale for social risk factors included? ☒Yes ☐No

13b. Are social risk factors included in risk model? ☒Yes ☐No

13c. Any concerns regarding the risk-adjustment approach?

***TIPS**: Consider the following: **If measure is risk adjusted**: If the developer asserts there is **no conceptual basis** for adjusting this measure for social risk factors, do you agree with the rationale? Are the candidate and final variables included in the risk adjustment model adequately described for the measure to be implemented? Are the candidate and final variables included in the risk adjustment model adequately described for the measure to be implemented? Are all of the risk adjustment variables present at the start of care (if not, do you agree with the rationale)? If social risk factors are not included in the risk-adjustment approach, do you agree with the developer's decision? Is an appropriate risk-adjustment strategy included in the measure (e.g., adequate model discrimination and calibration)? Are all statistical model specifications included, including a "clinical model only" if social risk factors are included in the final model? If a measure is NOT risk-adjusted, is a justification for **not risk adjusting** provided (conceptual and/or empirical)? Is there any evidence that contradicts the developer's rationale and analysis for not risk-adjusting?*

☐Yes (please explain below then go to Question #14)

☒No (go to Question #14)

☐Not applicable (e.g., this is a structure or process measure that is not risk-adjusted; go to Question #14)

68. Analysis of potential threats to validity: Any concerns regarding ability to identify meaningful differences in performance or overall poor performance?

REFERENCE: Testing attachment, section 2b4.

☐Yes (please explain below then go to Question #15)

☒No (go to Question #15)

69. Analysis of potential threats to validity: Any concerns regarding comparability of results if multiple data sources or methods are specified?

REFERENCE: Testing attachment, section 2b5.

☐Yes (please explain below then go to Question #16)

☒No (go to Question #16)

☐Not applicable (go to Question #16)

70. Analysis of potential threats to validity: Any concerns regarding missing data?

REFERENCE: Testing attachment, section 2b6.

☐Yes (please explain below then go to Question #17)

☒No (go to Question #17)

ASSESSMENT OF MEASURE TESTING

71. Was empirical validity testing conducted using the measure as specified and with appropriate statistical tests?

REFERENCE: Testing attachment, section 2b1.

***TIPS**: Answer no if: only face validity testing was performed; only refer to clinical evidence; only descriptive statistics; only describe process for data management/cleaning/computer programming; testing does not match measure specifications (i.e. data, eMeasure, level, setting, patients).*

☒Yes (go to Question #18)

☐No (please explain below, then skip Questions #18-23 and go to Question #24)

72. Was validity testing conducted with computed performance measure scores for each measured entity?

REFERENCE: Testing attachment, section 2b1.

***TIPS**: Answer no if: one overall score for all patients in sample used for testing patient-level data.*

☒Yes (go to Question #19)

☐No (please explain below, then skip questions #19-20 and go to Question #21)

73. Was the method described and appropriate for assessing conceptually and theoretically sound hypothesized relationships?

REFERENCE: Testing attachment, section 2b1.

***TIPS**: For example: correlation of the performance measure score on this measure and other performance measures; differences in performance scores between groups known to differ on quality; other accepted method with description of how it assesses validity of the performance score*

☒Yes (go to Question #20)

☐No (please explain below, then go to Question #20 and rate as INSUFFICIENT)

74. **RATING (measure score)** - Based on the measure score results (significance, strength) and scope of testing (number of measured entities and representativeness) and analysis of potential threats, what is the level of certainty or confidence that the performance measure scores are a valid indicator of quality?

☐High (go to Question #21)

☒Moderate (go to Question #21)

☐Low (please explain below then go to Question #21)

☐Insufficient (go to Question #21)

75. Was validity testing conducted with patient-level data elements?

REFERENCE: Testing attachment, section 2b1.

***TIPS**: Prior validity studies of the same data elements may be submitted*

☒Yes (go to Question #22)

☐No (if there is score-level testing that you rated something other than INSUFFICIENT in Question #20, skip questions #22-25, and go to Question #26 (OVERALL VALIDITY); otherwise, skip questions #22-23 and go to Question #24)

76. Was the method described and appropriate for assessing the accuracy of ALL critical data elements? *NOTE that data element validation from the literature is acceptable.*

***REFERENCE:** Testing attachment, section 2b1.*

***TIPS**: For example: Data validity/accuracy as compared to authoritative source- sensitivity, specificity, PPV, NPV; other accepted method with description of how it assesses validity of the data elements.*

Answer No if: only assessed percent agreement; did not assess separately for all data elements (at least numerator, denominator, exclusions)

☒Yes (go to Question #23)

☐No (please explain below, then go to Question #23 and rate as INSUFFICIENT)

77. **RATING (data element)** - Based on the data element testing results (significance, strength) and scope of testing (number and representativeness of patients and entities) and analysis of potential threats, what is the level of certainty or confidence that the data used in the measure are valid?

☒Moderate (skip Questions #24-25 and go to Question #26)

☐Low (please explain below, skip Questions #24-25 and go to Question #26)

☐Insufficient (go to Question #24 only if no other empirical validation was conducted OR if the measure has not been previously endorsed; otherwise, skip Questions #24-25 and go to Question #26)

78. Was face validity systematically assessed by recognized experts to determine agreement on whether the computed performance measure score from the measure as specified can be used to distinguish good and poor quality?

NOTE: If appropriate empirical testing has been conducted, then evaluation of face validity is not necessary; you should skip this question and Question 25, and answer Question #26 based on your answers to Questions #20 and/or #23]

REFERENCE: Testing attachment, section 2b1.

***TIPS**: Answer no if: focused on data element accuracy/availability/feasibility/other topics; the degree of consensus and any areas of disagreement not provided/discussed.*

☐Yes (go to Question #25)

☐No (please explain below, skip question #25, go to Question #26 (OVERALL VALIDITY) and rate as INSUFFICIENT)

79. **RATING (face validity)** - Do the face validity testing results indicate substantial agreement that the performance measure score from the measure as specified can be used to distinguish quality AND potential threats to validity are not a problem, OR are adequately addressed so results are not biased?

REFERENCE: Testing attachment, section 2b1.

***TIPS**: Face validity is no longer accepted for maintenance measures unless there is justification for why empirical validation is not possible and you agree with that justification.*

☐Yes (if a NEW measure, go to Question #26 (OVERALL VALIDITY) and rate as MODERATE)

☐Yes (if a MAINTENANCE measure, do you agree with the justification for not conducting empirical testing? If no, go to Question #26 (OVERALL VALIDITY) and rate as INSUFFICIENT; otherwise, rate Question #26 as MODERATE)

☐No (please explain below, go to Question #26 (OVERALL VALIDITY) and rate AS LOW)

OVERALL VALIDITY RATING

80. **OVERALL RATING OF VALIDITY** taking into account the results and scope of all testing and analysis of potential threats.

☐High (NOTE: Can be HIGH only if score-level testing has been conducted)

☒Moderate (NOTE: Moderate is the highest eligible rating if score-level testing has NOT been conducted)

☐Low (please explain below) [NOTE: Should rate LOW if you believe that there are threats to validity and/or threats to validity were not assessed]

☐Insufficient (if insufficient, please explain below) [NOTE: For instrument-based measures and some composite measures, testing at both the score level and the data element level is required; if not conducted, should rate as INSUFFICIENT—please check with NQF staff if you have questions.]

FOR COMPOSITE MEASURES ONLY: Empirical analyses to support composite construction

81. What is the level of certainty or confidence that the empirical analysis demonstrates that the component measures add value to the composite and that the aggregation and weighting rules are consistent with the quality construct?

REFERENCE: Testing attachment, section 2c

***TIPS**: Consider the following: Do the component measures fit the quality construct? Are the objectives of parsimony and simplicity achieved while supporting the quality construct?*

☐High

☐Moderate

☐Low (please explain below)

☐Insufficient (please explain below)

Review #4: Scientific Acceptability

Measure Number: 2473

Measure Title: Hybrid Hospital 30-Day All-Cause Risk-Standardized Mortality Rate Following AMI

Scientific Acceptability: Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. Measures must be judged to meet the subcriteria for both reliability and validity to pass this criterion

Instructions for filling out this form:

- Please complete this form for each measure you are evaluating.
- Please pay close attention to the skip logic directions. ***Directives that require you to skip questions are marked in red font.***
- If you are unable to check a box, please highlight or shade the box for your response.
- You must answer the "overall rating" item for both Reliability and Validity. Also, be sure to answer the composite measure question at the end of the form if your measure is a composite.
- For several questions, we have noted which sections of the submission documents you should ***REFERENCE*** and provided ***TIPS*** to help you answer them.
- ***It is critical that you explain your thinking/rationale if you check boxes that require an explanation.*** Please add your explanation directly below the checkbox in a different font color***.*** Also, feel free to add additional explanation, even if you select a checkbox where an explanation is not requested (if you do so, please type this text directly below the appropriate checkbox).
- ***Please refer to the*** Measure Evaluation Criteria and Guidance document ***(pages 18-24) and the 2-page*** Key Points document ***when evaluating your measures***. This evaluation form is an adaptation of Alogorithms 2 and 3, which provide guidance on rating the Reliability and Validity subcriteria.
- ***Remember*** that testing at either the data element level **OR** the measure score level is accepted for some types of measures, but not all (e.g., instrument-based measures, composite measures), and therefore, the embedded rating instructions may not be appropriate for all measures.
- ***Please base your evaluations solely on the submission materials provided by developers.*** NQF strongly discourages the use of outside articles or other resources, even if they are cited in the submission materials. If you require further information or clarification to conduct your evaluation, please communicate with NQF staff (methodspanel@qualityforum.org).

RELIABILITY

82. Are submitted specifications precise, unambiguous, and complete so that they can be consistently implemented?

    REFERENCE: "MIF_xxxx" document

    NOTE*: NQF staff will conduct a separate, more technical, check of eMeasure (eCQM) specifications, value sets, logic, and feasibility, so no need to consider these in your evaluation.*

    TIPS*: Consider the following: Are all the data elements clearly defined? Are all appropriate codes included? Is the logic or calculation algorithm clear? Is it likely this measure can be consistently implemented?*

    ☒Yes (go to Question #2)

    ☐No (please explain below, and go to Question #2) NOTE that even though ***non-precise specifications should result in an overall LOW rating for reliability***, we still want you to look at the testing results.

83. Was empirical reliability testing (at the data element or measure score level) conducted using statistical tests with the measure as specified?

    REFERENCE: "MIF_xxxx" document for specifications, testing attachment questions 1.1-1.4 and section 2a2

***TIPS**: Check the "NO" box below if: only descriptive statistics are provided; only describes process for data management/cleaning/computer programming; testing does not match measure specifications (i.e., data source, level of analysis, included patients, etc.)*

☒Yes (go to Question #3)

☐No, there is reliability testing information, but *not* using statistical tests and/or not for the measure as specified **OR** there is no reliability testing (please explain below, skip Questions #3-8, then go to Question #9)

84. Was reliability testing conducted with computed performance measure scores for each measured entity?

REFERENCE: "Testing attachment_xxx", section 2a2.1 and 2a2.2

***TIPS**: Answer no if: only one overall score for all patients in sample used for testing patient-level data*

☒Yes (go to Question #4)

☐No (skip Questions #4-5 and go to Question #6)

85. Was the method described and appropriate for assessing the proportion of variability due to real differences among measured entities? *NOTE: If multiple methods used, at least one must be appropriate.*

REFERENCE: Testing attachment, section 2a2.2

***TIPS**: Examples of appropriate methods include signal-to-noise analysis (e.g. Adams/RAND tutorial); random split-half correlation; other accepted method with description of how it assesses reliability of the performance score.*

☒Yes (go to Question #5)

☐No (please explain below, then go to question #5 and rate as INSUFFICIENT)

86. **RATING (score level)** - What is the level of certainty or confidence that the performance measure scores are reliable?

REFERENCE: Testing attachment, section 2a2.2

***TIPS**: Consider the following: Is the test sample adequate to generalize for widespread implementation? Do the results demonstrate sufficient reliability so that differences in performance can be identified?*

☐High (go to Question #6)

☒Moderate (go to Question #6)

Please note that the estimated reliability score of 0.42 is in the low end of the moderately reliable category as defined by Landis & Koch, 1977.

☐Low (please explain below then go to Question #6)

☐Insufficient (go to Question #6)

87. Was reliability testing conducted with patient-level data elements that are used to construct the performance measure?

REFERENCE: Testing attachment, section 2a2.

***TIPS**: Prior reliability studies of the same data elements may be submitted; if comparing abstraction to "authoritative source/gold standard" go to Question #9)*

☐Yes (go to Question #7)

☒No (if there is score-level testing that you rated something other than INSUFFICIENT in Question #5, skip questions #7-9, then go to Question #10 (OVERALL RELIABILITY); otherwise, skip questions #7-8 and go to Question #9)

88. Was the method described and appropriate for assessing the reliability of ALL critical data elements?

REFERENCE: Testing attachment, section 2a2.2

***TIPS**: For example: inter-abstractor agreement (ICC, Kappa); other accepted method with description of how it assesses reliability of the data elements*

Answer no if: only assessed percent agreement; did not assess separately for all data elements (at least numerator, denominator, exclusions)

☐Yes (go to Question #8)

☐No (if no, please explain below, then go to Question #8 and rate as INSUFFICIENT)

89. **RATING (data element)** – Based on the reliability statistic and scope of testing (number and representativeness of patients and entities), what is the level of certainty or confidence that the data used in the measure are reliable?

REFERENCE: Testing attachment, section 2a2

***TIPS**: Consider the following: Is the test sample adequate to generalize for widespread implementation? Can data elements be collected consistently?*

☐Moderate (skip Question #9 and go to Question #10 (OVERALL RELIABILITY); if score-level testing was NOT conducted, rate Question #10 as MODERATE)

☐Low (skip Question #9 and go to Question #10 (OVERALL RELIABILITY); if score-level testing was NOT conducted, rate Question #10 (OVERALL RELIABILITY) as LOW)

☐Insufficient (go to Question #9)

90. Was **empirical VALIDITY testing** of patient-level data conducted?

REFERENCE: testing attachment section 2b1.

NOTE: Skip this question if empirical reliability testing was conducted and you have rated Question #5 and/or #8 as anything other than INSUFFICIENT)

TIP:** You should answer this question ONLY if score-level or data element reliability testing was NOT conducted or if the methods used were NOT appropriate. For most measures, NQF will accept data element validity testing in lieu of reliability testing—but **check with NQF staff before proceeding, to verify.

☐Yes (go to Question #10 and answer using your rating from data element validity testing – Question #23)

☐No (please explain below, go to Question #10 (OVERALL RELIABILITY) and rate it as INSUFFICIENT. Then go to Question #11.)

OVERALL RELIABILITY RATING

91. **OVERALL RATING OF RELIABILITY** taking into account precision of specifications (see Question #1) and all testing results:

☐High (NOTE: Can be HIGH only if score-level testing has been conducted)

☒Moderate (NOTE: Moderate is the highest eligible rating if score-level testing has not been conducted)

☐Low (please explain below) [NOTE: Should rate LOW if you believe specifications are NOT precise, unambiguous, and complete]

☐Insufficient (please explain below) [NOTE: For most measure types, testing at both the score level and the data element level is not required, but check with NQF staff]

VALIDITY

ASSESSMENT OF THREATS TO VALIDITY

92. Were potential threats to validity that are relevant to the measure empirically assessed ()?

REFERENCE: Testing attachment, section 2b2-2b6

***TIPS**: Threats to validity that should be assessed include: exclusions; need for risk adjustment; ability to identify statistically significant and meaningful differences; multiple sets of specifications; missing data/nonresponse.*

☒Yes (go to Question #12)

☐No (please explain below and then go to Question #12) [NOTE that ***non-assessment of applicable threats should result in an overall INSUFFICENT rating for validity***]

93. Analysis of potential threats to validity: Any concerns with measure exclusions?

REFERENCE: Testing attachment, section 2b2.

***TIPS**: Consider the following: Are the exclusions consistent with the evidence? Are any patients or patient groups inappropriately excluded from the measure? Are the exclusions/exceptions of sufficient frequency and variation across providers to be needed (and outweigh the data collection burden)? If patient preference (e.g., informed decisionmaking) is a basis for exclusion, does it impact performance and if yes, is the measure specified so that the information about patient preference and the effect on the measure is transparent?*

☐Yes (please explain below then go to Question #13)

☒No (go to Question #13)

☐Not applicable (i.e., there are no exclusions specified for the measure; go to Question #13)

94. Analysis of potential threats to validity: Risk-adjustment (this applies to <u>all</u> outcome, cost, and resource use measures and "NOT APPLICABLE" is not an option for those measures; the risk-adjustment questions (13a-13c, below) also may apply to other types of measures)

REFERENCE: Testing attachment, section 2b3.

13a. Is a conceptual rationale for social risk factors included? ☒Yes ☐No

13b. Are social risk factors included in risk model? ☐Yes ☒No

13c. Any concerns regarding the risk-adjustment approach?

***TIPS**: Consider the following: **If measure is risk adjusted**: If the developer asserts there is **no conceptual basis** for adjusting this measure for social risk factors, do you agree with the rationale? Are the candidate and final variables included in the risk adjustment model adequately described for the measure to be implemented? Are the candidate and final variables included in the risk adjustment model adequately described for the measure to be implemented? Are all of the risk adjustment variables present at the start of care (if not, do you agree with the rationale)? If social risk factors are not included in the risk-adjustment approach, do you agree with the developer's decision? Is an appropriate risk-adjustment strategy included in the measure (e.g., adequate model discrimination and calibration)? Are all statistical model specifications included, including a "clinical model only" if social risk factors are included in the final model? If a measure is NOT risk-adjusted, is a justification for **not risk adjusting** provided (conceptual and/or empirical)? Is there any evidence that contradicts the developer's rationale and analysis for not risk-adjusting?*

☐Yes (please explain below then go to Question #14)

☒No (go to Question #14)

☐Not applicable (e.g., this is a structure or process measure that is not risk-adjusted; go to Question #14)

95. Analysis of potential threats to validity: Any concerns regarding ability to identify meaningful differences in performance or overall poor performance?

REFERENCE: Testing attachment, section 2b4.

☐Yes (please explain below then go to Question #15)

☒No (go to Question #15)

96. Analysis of potential threats to validity: Any concerns regarding comparability of results if multiple data sources or methods are specified?

REFERENCE: Testing attachment, section 2b5.

☐Yes (please explain below then go to Question #16)

☒No (go to Question #16)

☐Not applicable (go to Question #16)

97. Analysis of potential threats to validity: Any concerns regarding missing data?

REFERENCE: Testing attachment, section 2b6.

☐Yes (please explain below then go to Question #17)

☒No (go to Question #17)

ASSESSMENT OF MEASURE TESTING

98. Was empirical validity testing conducted using the measure as specified and with appropriate statistical tests?

REFERENCE: Testing attachment, section 2b1.

***TIPS**: Answer no if: only face validity testing was performed; only refer to clinical evidence; only descriptive statistics; only describe process for data management/cleaning/computer programming; testing does not match measure specifications (i.e. data, eMeasure, level, setting, patients).*

☒Yes (go to Question #18)

☐No (please explain below, then skip Questions #18-23 and go to Question #24)

99. Was validity testing conducted with computed performance measure scores for each measured entity?

REFERENCE: Testing attachment, section 2b1.

***TIPS**: Answer no if: one overall score for all patients in sample used for testing patient-level data.*

☒Yes (go to Question #19)

☐No (please explain below, then skip questions #19-20 and go to Question #21)

100. Was the method described and appropriate for assessing conceptually and theoretically sound hypothesized relationships?

REFERENCE: Testing attachment, section 2b1.

***TIPS**: For example: correlation of the performance measure score on this measure and other performance measures; differences in performance scores between groups known to differ on quality; other accepted method with description of how it assesses validity of the performance score*

☒Yes (go to Question #20)

☐No (please explain below, then go to Question #20 and rate as INSUFFICIENT)

101. **RATING (measure score)** - Based on the measure score results (significance, strength) and scope of testing (number of measured entities and representativeness) and analysis of potential threats, what is the level of certainty or confidence that the performance measure scores are a valid indicator of quality?

☒High (go to Question #21)

☐Moderate (go to Question #21)

☐Low (please explain below then go to Question #21)

☐Insufficient (go to Question #21)

102. Was validity testing conducted with patient-level data elements?

REFERENCE: Testing attachment, section 2b1.

***TIPS**: Prior validity studies of the same data elements may be submitted*

☒Yes (go to Question #22)

☐No (if there is score-level testing that you rated something other than INSUFFICIENT in Question #20, skip questions #22-25, and go to Question #26 (OVERALL VALIDITY); otherwise, skip questions #22-23 and go to Question #24)

103. Was the method described and appropriate for assessing the accuracy of ALL critical data elements? *NOTE that data element validation from the literature is acceptable.*

***REFERENCE:** Testing attachment, section 2b1.*

***TIPS**: For example: Data validity/accuracy as compared to authoritative source- sensitivity, specificity, PPV, NPV; other accepted method with description of how it assesses validity of the data elements.*

Answer No if: only assessed percent agreement; did not assess separately for all data elements (at least numerator, denominator, exclusions)

☒Yes (go to Question #23)

☐No (please explain below, then go to Question #23 and rate as INSUFFICIENT)

104. **RATING (data element)** - Based on the data element testing results (significance, strength) and scope of testing (number and representativeness of patients and entities) and analysis of potential threats, what is the level of certainty or confidence that the data used in the measure are valid?

☒Moderate (skip Questions #24-25 and go to Question #26)

☐Low (please explain below, skip Questions #24-25 and go to Question #26)

☐Insufficient (go to Question #24 only if no other empirical validation was conducted OR if the measure has not been previously endorsed; otherwise, skip Questions #24-25 and go to Question #26)

105. Was face validity systematically assessed by recognized experts to determine agreement on whether the computed performance measure score from the measure as specified can be used to distinguish good and poor quality?

NOTE: If appropriate empirical testing has been conducted, then evaluation of face validity is not necessary; you should skip this question and Question 25, and answer Question #26 based on your answers to Questions #20 and/or #23]

REFERENCE: Testing attachment, section 2b1.

***TIPS**: Answer no if: focused on data element accuracy/availability/feasibility/other topics; the degree of consensus and any areas of disagreement not provided/discussed.*

☐Yes (go to Question #25)

☐No (please explain below, skip question #25, go to Question #26 (OVERALL VALIDITY) and rate as INSUFFICIENT)

106. **RATING (face validity)** - Do the face validity testing results indicate substantial agreement that the performance measure score from the measure as specified can be used to distinguish quality AND potential threats to validity are not a problem, OR are adequately addressed so results are not biased?

REFERENCE: Testing attachment, section 2b1.

***TIPS**: Face validity is no longer accepted for maintenance measures unless there is justification for why empirical validation is not possible and you agree with that justification.*

☐Yes (if a NEW measure, go to Question #26 (OVERALL VALIDITY) and rate as MODERATE)

☐Yes (if a MAINTENANCE measure, do you agree with the justification for not conducting empirical testing? If no, go to Question #26 (OVERALL VALIDITY) and rate as INSUFFICIENT; otherwise, rate Question #26 as MODERATE)

☐No (please explain below, go to Question #26 (OVERALL VALIDITY) and rate AS LOW)

OVERALL VALIDITY RATING

107. **OVERALL RATING OF VALIDITY** taking into account the results and scope of all testing and analysis of potential threats.

☒High (NOTE: Can be HIGH only if score-level testing has been conducted)

☐Moderate (NOTE: Moderate is the highest eligible rating if score-level testing has NOT been conducted)

☐Low (please explain below) [NOTE: Should rate LOW if you believe that there are threats to validity and/or threats to validity were not assessed]

☐Insufficient (if insufficient, please explain below) [NOTE: For instrument-based measures and some composite measures, testing at both the score level and the data element level is required; if not conducted, should rate as INSUFFICIENT—please check with NQF staff if you have questions.]

FOR COMPOSITE MEASURES ONLY: Empirical analyses to support composite construction

108. What is the level of certainty or confidence that the empirical analysis demonstrates that the component measures add value to the composite and that the aggregation and weighting rules are consistent with the quality construct?

REFERENCE: Testing attachment, section 2c

***TIPS**: Consider the following: Do the component measures fit the quality construct? Are the objectives of parsimony and simplicity achieved while supporting the quality construct?*

☐High

☐Moderate

☐Low (please explain below)

☐Insufficient (please explain below)

Review #5: Scientific Acceptability

Measure Number: 2473

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

Scientific Acceptability: Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. Measures must be judged to meet the subcriteria for both reliability and validity to pass this criterion

Instructions for filling out this form:

- Please complete this form for each measure you are evaluating.
- Please pay close attention to the skip logic directions. ***Directives that require you to skip questions are marked in red font.***
- If you are unable to check a box, please highlight or shade the box for your response.
- You must answer the "overall rating" item for both Reliability and Validity. Also, be sure to answer the composite measure question at the end of the form if your measure is a composite.
- For several questions, we have noted which sections of the submission documents you should ***REFERENCE*** and provided ***TIPS*** to help you answer them.
- ***It is critical that you explain your thinking/rationale if you check boxes that require an explanation.*** Please add your explanation directly below the checkbox in a different font color***.*** Also, feel free to add additional explanation, even if you select a checkbox where an explanation is not requested (if you do so, please type this text directly below the appropriate checkbox).
- ***Please refer to the*** Measure Evaluation Criteria and Guidance document ***(pages 18-24) and the 2-page*** Key Points document ***when evaluating your measures***. This evaluation form is an adaptation of Alogorithms 2 and 3, which provide guidance on rating the Reliability and Validity subcriteria.
- ***Remember*** that testing at either the data element level **OR** the measure score level is accepted for some types of measures, but not all (e.g., instrument-based measures, composite measures), and therefore, the embedded rating instructions may not be appropriate for all measures.
- ***Please base your evaluations solely on the submission materials provided by developers.*** NQF strongly discourages the use of outside articles or other resources, even if they are cited in the submission materials. If you require further information or clarification to conduct your evaluation, please communicate with NQF staff (methodspanel@qualityforum.org).

RELIABILITY

109. Are submitted specifications precise, unambiguous, and complete so that they can be consistently implemented?

REFERENCE: "MIF_xxxx" document

NOTE*: NQF staff will conduct a separate, more technical, check of eMeasure (eCQM) specifications, value sets, logic, and feasibility, so no need to consider these in your evaluation.*

TIPS*: Consider the following: Are all the data elements clearly defined? Are all appropriate codes included? Is the logic or calculation algorithm clear? Is it likely this measure can be consistently implemented?*

☒ Yes (go to Question #2)

☐ No (please explain below, and go to Question #2) NOTE that even though ***non-precise specifications should result in an overall LOW rating for reliability***, we still want you to look at the testing results.

110. Was empirical reliability testing (at the data element or measure score level) conducted using statistical tests with the measure as specified?

REFERENCE: "MIF_xxxx" document for specifications, testing attachment questions 1.1-1.4 and section 2a2

***TIPS**: Check the "NO" box below if: only descriptive statistics are provided; only describes process for data management/cleaning/computer programming; testing does not match measure specifications (i.e., data source, level of analysis, included patients, etc.)*

☒Yes (go to Question #3)

☐No, there is reliability testing information, but *not* using statistical tests and/or not for the measure as specified **OR** there is no reliability testing (please explain below, skip Questions #3-8, then go to Question #9)

111. Was reliability testing conducted with computed performance measure scores for each measured entity?

REFERENCE: "Testing attachment_xxx", section 2a2.1 and 2a2.2

***TIPS**: Answer no if: only one overall score for all patients in sample used for testing patient-level data*

☒Yes (go to Question #4)

☐No (skip Questions #4-5 and go to Question #6)

112. Was the method described and appropriate for assessing the proportion of variability due to real differences among measured entities? *NOTE: If multiple methods used, at least one must be appropriate.*

REFERENCE: Testing attachment, section 2a2.2

***TIPS**: Examples of appropriate methods include signal-to-noise analysis (e.g. Adams/RAND tutorial); random split-half correlation; other accepted method with description of how it assesses reliability of the performance score.*

☒Yes (go to Question #5)

☒No (please explain below, then go to question #5 and rate as INSUFFICIENT)

113. **RATING (score level)** - What is the level of certainty or confidence that the performance measure scores are reliable?

REFERENCE: Testing attachment, section 2a2.2

***TIPS**: Consider the following: Is the test sample adequate to generalize for widespread implementation? Do the results demonstrate sufficient reliability so that differences in performance can be identified?*

☐High (go to Question #6)

☒Moderate (go to Question #6)

☐Low (please explain below then go to Question #6)

☐Insufficient (go to Question #6)

114. Was reliability testing conducted with patient-level data elements that are used to construct the performance measure?

REFERENCE: Testing attachment, section 2a2.

***TIPS**: Prior reliability studies of the same data elements may be submitted; if comparing abstraction to "authoritative source/gold standard" go to Question #9)*

☐Yes (go to Question #7)

☒No (if there is score-level testing that you rated something other than INSUFFICIENT in Question #5, skip questions #7-9, then go to Question #10 (OVERALL RELIABILITY); otherwise, skip questions #7-8 and go to Question #9)

115. Was the method described and appropriate for assessing the reliability of ALL critical data elements?

REFERENCE: Testing attachment, section 2a2.2

***TIPS**: For example: inter-abstractor agreement (ICC, Kappa); other accepted method with description of how it assesses reliability of the data elements*

Answer no if: only assessed percent agreement; did not assess separately for all data elements (at least numerator, denominator, exclusions)

☐Yes (go to Question #8)

☐No (if no, please explain below, then go to Question #8 and rate as INSUFFICIENT)

116. **RATING (data element)** – Based on the reliability statistic and scope of testing (number and representativeness of patients and entities), what is the level of certainty or confidence that the data used in the measure are reliable?

REFERENCE: Testing attachment, section 2a2

***TIPS**: Consider the following: Is the test sample adequate to generalize for widespread implementation? Can data elements be collected consistently?*

☐Moderate (skip Question #9 and go to Question #10 (OVERALL RELIABILITY); if score-level testing was NOT conducted, rate Question #10 as MODERATE)

☐Low (skip Question #9 and go to Question #10 (OVERALL RELIABILITY); if score-level testing was NOT conducted, rate Question #10 (OVERALL RELIABILITY) as LOW)

☐Insufficient (go to Question #9)

117. Was **empirical VALIDITY testing** of patient-level data conducted?

REFERENCE: testing attachment section 2b1.

NOTE: Skip this question if empirical reliability testing was conducted and you have rated Question #5 and/or #8 as anything other than INSUFFICIENT)

TIP:** You should answer this question ONLY if score-level or data element reliability testing was NOT conducted or if the methods used were NOT appropriate. For most measures, NQF will accept data element validity testing in lieu of reliability testing—but **check with NQF staff before proceeding, to verify.

☐Yes (go to Question #10 and answer using your rating from data element validity testing – Question #23)

☐No (please explain below, go to Question #10 (OVERALL RELIABILITY) and rate it as INSUFFICIENT. Then go to Question #11.)

OVERALL RELIABILITY RATING

118. **OVERALL RATING OF RELIABILITY** taking into account precision of specifications (see Question #1) and all testing results:

☐High (NOTE: Can be HIGH only if score-level testing has been conducted)

☒Moderate (NOTE: Moderate is the highest eligible rating if score-level testing has not been conducted)

☐Low (please explain below) [NOTE: Should rate LOW if you believe specifications are NOT precise, unambiguous, and complete]

☐Insufficient (please explain below) [NOTE: For most measure types, testing at both the score level and the data element level is not required, but check with NQF staff]

VALIDITY

ASSESSMENT OF THREATS TO VALIDITY

119. Were potential threats to validity that are relevant to the measure empirically assessed ()?

REFERENCE: Testing attachment, section 2b2-2b6

***TIPS**: Threats to validity that should be assessed include: exclusions; need for risk adjustment; ability to identify statistically significant and meaningful differences; multiple sets of specifications; missing data/nonresponse.*

☐Yes (go to Question #12)

☒No (please explain below and then go to Question #12) [NOTE that ***non-assessment of applicable threats should result in an overall INSUFFICENT rating for validity***]

A major threat to validity is that there is no mention of disparate data, missing data, outlier data, and erroneous data due to events like a hemolyzed lab specimen.

120. Analysis of potential threats to validity: Any concerns with measure exclusions?

REFERENCE: Testing attachment, section 2b2.

***TIPS**: Consider the following: Are the exclusions consistent with the evidence? Are any patients or patient groups inappropriately excluded from the measure? Are the exclusions/exceptions of sufficient frequency and variation across providers to be needed (and outweigh the data collection burden)? If patient preference (e.g., informed decisionmaking) is a basis for exclusion, does it impact performance and if yes, is the measure specified so that the information about patient preference and the effect on the measure is transparent?*

☒Yes (please explain below then go to Question #13)

I agree with the exclusion criteria but I am very concerned with the very low frequency for the size of the data set (217,723 admissions). For example, the percentage of Discharged Against Medical Advice is reported at 0.24%. That is significantly lower than the more typical rate reported around 2%. In the AHRQ Healthcare Cost and Utilization Project- Trends in Emergency Department Visits, the report showed a national rate of 1.8% for 2014. I would recommend that the measure developers address these threats to validity before endorsing this for a payment program.

☐No (go to Question #13)

☐Not applicable (i.e., there are no exclusions specified for the measure; go to Question #13)

121. Analysis of potential threats to validity: Risk-adjustment (this applies to <u>all</u> outcome, cost, and resource use measures and "NOT APPLICABLE" is not an option for those measures; the risk-adjustment questions (13a-13c, below) also may apply to other types of measures)

REFERENCE: Testing attachment, section 2b3.

13a. Is a conceptual rationale for social risk factors included? ☒Yes ☐No

13b. Are social risk factors included in risk model? ☒Yes ☐No

13c. Any concerns regarding the risk-adjustment approach?

***TIPS**: Consider the following: **If measure is risk adjusted**: If the developer asserts there is **no conceptual basis** for adjusting this measure for social risk factors, do you agree with the rationale? Are the candidate and final variables included in the risk adjustment model adequately described for the measure to be implemented? Are the candidate and final variables included in the risk adjustment model adequately described for the measure to be implemented? Are all of the risk adjustment variables present at the start of care (if not, do you agree with the rationale)? If social risk factors are not included in the risk-adjustment approach, do you agree with the developer's decision? Is an appropriate risk-adjustment strategy included in the measure (e.g., adequate model discrimination and calibration)? Are all statistical model specifications included, including a "clinical model only" if social risk factors are included in the final model? If a measure is NOT risk-adjusted, is a justification for **not risk adjusting** provided (conceptual and/or empirical)? Is there any evidence that contradicts the developer's rationale and analysis for not risk-adjusting?*

☐Yes (please explain below then go to Question #14)

☒No (go to Question #14)

☐Not applicable (e.g., this is a structure or process measure that is not risk-adjusted; go to Question #14)

122. Analysis of potential threats to validity: Any concerns regarding ability to identify meaningful differences in performance or overall poor performance?

REFERENCE: Testing attachment, section 2b4.

☒Yes (please explain below then go to Question #15)

As noted by the measure developers in section 2b4; no method for discriminating hospital performance has been determined. CMS only uses "better than" or "worse than" the national rates for publicly reporting. The overall statistical analysis does support that the model can discern meaningful differences but not what those differences mean. It would be helpful to do a predictive mortality analysis or observed-to-expected mortality rate to give meaning to the differences in hospitals.

☐No (go to Question #15)

123. Analysis of potential threats to validity: Any concerns regarding comparability of results if multiple data sources or methods are specified?

REFERENCE: Testing attachment, section 2b5.

☐Yes (please explain below then go to Question #16)

☐No (go to Question #16)

☒Not applicable (go to Question #16)

124. Analysis of potential threats to validity: Any concerns regarding missing data?

REFERENCE: Testing attachment, section 2b6.

☐Yes (please explain below then go to Question #17)

☒No (go to Question #17)

ASSESSMENT OF MEASURE TESTING

125. Was empirical validity testing conducted using the measure as specified and with appropriate statistical tests?

REFERENCE: Testing attachment, section 2b1.

***TIPS**: Answer no if: only face validity testing was performed; only refer to clinical evidence; only descriptive statistics; only describe process for data management/cleaning/computer programming; testing does not match measure specifications (i.e. data, eMeasure, level, setting, patients).*

☒Yes (go to Question #18)

☐No (please explain below, then skip Questions #18-23 and go to Question #24)

126. Was validity testing conducted with computed performance measure scores for each measured entity?

REFERENCE: Testing attachment, section 2b1.

***TIPS**: Answer no if: one overall score for all patients in sample used for testing patient-level data.*

☒Yes (go to Question #19)

☐No (please explain below, then skip questions #19-20 and go to Question #21)

127. Was the method described and appropriate for assessing conceptually and theoretically sound hypothesized relationships?

REFERENCE: Testing attachment, section 2b1.

***TIPS**: For example: correlation of the performance measure score on this measure and other performance measures; differences in performance scores between groups known to differ on quality; other accepted method with description of how it assesses validity of the performance score*

☒Yes (go to Question #20)

☐No (please explain below, then go to Question #20 and rate as INSUFFICIENT)

128. **RATING (measure score)** - Based on the measure score results (significance, strength) and scope of testing (number of measured entities and representativeness) and analysis of potential threats, what is the level of certainty or confidence that the performance measure scores are a valid indicator of quality?

☐High (go to Question #21)

☒Moderate (go to Question #21)

The model only supports that it can measure a difference between entities but does not support that those differences indicate better or worse quality of care.

☐Low (please explain below then go to Question #21)

☐Insufficient (go to Question #21)

129. Was validity testing conducted with patient-level data elements?

REFERENCE: Testing attachment, section 2b1.

***TIPS**: Prior validity studies of the same data elements may be submitted*

☒Yes (go to Question #22)

☐No (if there is score-level testing that you rated something other than INSUFFICIENT in Question #20, skip questions #22-25, and go to Question #26 (OVERALL VALIDITY); otherwise, skip questions #22-23 and go to Question #24)

130. Was the method described and appropriate for assessing the accuracy of ALL critical data elements? *NOTE that data element validation from the literature is acceptable.*

***REFERENCE:** Testing attachment, section 2b1.*

***TIPS**: For example: Data validity/accuracy as compared to authoritative source- sensitivity, specificity, PPV, NPV; other accepted method with description of how it assesses validity of the data elements.*

Answer No if: only assessed percent agreement; did not assess separately for all data elements (at least numerator, denominator, exclusions)

☒Yes (go to Question #23)

☐No (please explain below, then go to Question #23 and rate as INSUFFICIENT)

131. **RATING (data element)** - Based on the data element testing results (significance, strength) and scope of testing (number and representativeness of patients and entities) and analysis of potential threats, what is the level of certainty or confidence that the data used in the measure are valid?

☒Moderate (skip Questions #24-25 and go to Question #26)

☐Low (please explain below, skip Questions #24-25 and go to Question #26)

☐Insufficient (go to Question #24 only if no other empirical validation was conducted OR if the measure has not been previously endorsed; otherwise, skip Questions #24-25 and go to Question #26)

132. Was face validity systematically assessed by recognized experts to determine agreement on whether the computed performance measure score from the measure as specified can be used to distinguish good and poor quality?

NOTE: If appropriate empirical testing has been conducted, then evaluation of face validity is not necessary; you should skip this question and Question 25, and answer Question #26 based on your answers to Questions #20 and/or #23]

REFERENCE: Testing attachment, section 2b1.

***TIPS**: Answer no if: focused on data element accuracy/availability/feasibility/other topics; the degree of consensus and any areas of disagreement not provided/discussed.*

☐Yes (go to Question #25)

☐No (please explain below, skip question #25, go to Question #26 (OVERALL VALIDITY) and rate as INSUFFICIENT)

133. **RATING (face validity)** - Do the face validity testing results indicate substantial agreement that the performance measure score from the measure as specified can be used to distinguish quality AND potential threats to validity are not a problem, OR are adequately addressed so results are not biased?

REFERENCE: Testing attachment, section 2b1.

***TIPS**: Face validity is no longer accepted for maintenance measures unless there is justification for why empirical validation is not possible and you agree with that justification.*

☐Yes (if a NEW measure, go to Question #26 (OVERALL VALIDITY) and rate as MODERATE)

☐Yes (if a MAINTENANCE measure, do you agree with the justification for not conducting empirical testing? If no, go to Question #26 (OVERALL VALIDITY) and rate as INSUFFICIENT; otherwise, rate Question #26 as MODERATE)

☐No (please explain below, go to Question #26 (OVERALL VALIDITY) and rate AS LOW)

OVERALL VALIDITY RATING

134. **OVERALL RATING OF VALIDITY** taking into account the results and scope of all testing and analysis of potential threats.

☐High (NOTE: Can be HIGH only if score-level testing has been conducted)

☐Moderate (NOTE: Moderate is the highest eligible rating if score-level testing has NOT been conducted)

☒Low (please explain below) [NOTE: Should rate LOW if you believe that there are threats to validity and/or threats to validity were not assessed]

See previous comments under threats to validity. I would like the measure developers to address these concerns.

☐Insufficient (if insufficient, please explain below) [NOTE: For instrument-based measures and some composite measures, testing at both the score level and the data element level is required; if not conducted, should rate as INSUFFICIENT—please check with NQF staff if you have questions.]

FOR COMPOSITE MEASURES ONLY: Empirical analyses to support composite construction

135. What is the level of certainty or confidence that the empirical analysis demonstrates that the component measures add value to the composite and that the aggregation and weighting rules are consistent with the quality construct?

REFERENCE: Testing attachment, section 2c

***TIPS**: Consider the following: Do the component measures fit the quality construct? Are the objectives of parsimony and simplicity achieved while supporting the quality construct?*

☐High

☐Moderate

☐Low (please explain below)

☐Insufficient (please explain below)

Committee Pre-evaluation Comments: Criteria 2: Scientific Acceptability of Measure Properties (including all 2a, 2b, and 2c)

2a1. Specifications:

- One step that may not have much impact but introduces a slight potential bias is: For the three-year combined data, when index admissions occur during the transition between measure reporting periods (June and July of each year) and both are randomly selected for inclusion in the measure, the measure includes only the June admission. July admissions are excluded to avoid assigning a single death to two admissions. This step would place all such patients in the denominator with reduced chance of appearing in the numerator of the measure, as the patient would have lived to at least until July.
- Reliability of the measure score by test-retest was 0.42 "moderate". This was at the level of the score."
- Mix of claims data and EHR data appears to have unaddressed issues.
- Data elements extracted from EMR are likely highly variable throughout patient's hospital stay (BP, cr, troponin)

This likely introduces significant error in adjustment and reliability.

- Overall reliability is listed as moderate. There is no patient level data testing, and moderate performance measure scores.
- This is measure is supposed to be use a hybrid of data collected from claims and from the EHR. However, the measure developers were unable to obtain hospital EHR records for this purpose and instead made use of registry data for BP, HR, troponin, creatinine (age is also obtained from claims). Medicare Advantage enrollment does not appear to be an exclusion criterion so it appears that variables derived from part A or part B claims were not actually tested as risk adjustors, although the write-up is somewhat opaque in this regard.
- Parsimonious set of risk adjustment variables but reasonable C-statistic. Multiple datasets used for testing reliability. Moderate reliability scores between development samples and validation samples.

2a2. Reliability testing:

- Apparently both the reliability and validity testing was not done with data extracted from the EHR. From the 2nd reviewer: “I’m confused by the use of registry data to validate the measure and then EMR data to validate the data elements. This may be all that is currently feasible, but it seems important to use the actual data sources from which the measure would be calculated to assess validity. Also very odd that they are using 9-year-old data when much more updated data (registry, EMR, and claims) certainly exist.” The reviewer also noted “High proportion of missing troponin levels, which may be salient given the small number of risk adjusters included in the clinical data”. The third reviewer mentioned adding some exclusions so the measure would be more specific to AMI care.
- No

- Yes, based on limited EHR testing (only 2 systems) and lack of detail on reconciling claims and EHR data.
- Because the reliability is moderate, the results should be reported with a calculated confidence interval. this is inherent an empiric population based measure.
- Cannot assess

2b1. Validity testing & 2b4-7. Threats to Validity (Statistically Significant Differences, Multiple Data Sources, Missing Data):

- The Methods Panel reviewed the measure for validity and were concerned that registry data were used to validate the measure score and EMR data were used to validate the data elements. I agree that both reliability and validity are moderate.
- Mixing claims and EHR data sources without sufficient reconciliation strategy explanation makes validity dubious.
- It is unclear what this measure adds to the NCDR measure that looks into the same hospital mortality with more data elements for adjustment. The CMS measure has a great strength that it has the full nominator (death) and denominator (all claims received). However, compared to the NCDR measure the adjustment and data collection is rather crude and rudimentary. Both measures compete in the same area - neither is adequately strong to be used for payment or public reporting.
- Because this measure does not have an absolute accuracy the impact of missing data is not as significant as it would be in a more precise measure.
- The elephant on the chest here is that as defined, cases will consist of an admixture of patients with STEMI, other ACS, and patients given an MI diagnosis because they had circulating troponin for non-coronary-related reasons. This mix would be expected to vary substantially from hospital to hospital, depending on coding culture, on whether they are a STEMI center, on how many STEMI centers in the patient catchment area, etc., etc. And none of that is accounted for in the risk adjustment.

Other:

- In the technical report, table 10, it is unclear why two coefficients are reported for two possible values of the dichotomous HR variable (HR<70 and HR>=70)
- Patient population in the registry is more selected than the population defined by inclusion/exclusion criteria because registry inclusion is voluntary AND requires the following conditions which are NOT part of the measure inclusion/exclusion criteria.
- 1) Ischemic symptoms at rest, lasting ≥10 minutes, occurring in the 24 hours before admission, or up to 72 hours for ST segment elevated myocardial infarction (STEMI); or
- 2) Electrocardiogram (ECG) changes associated with STEMI (new left bundle-branch block [LBBB] or persistent STEMI ≥1 mm in two or more contiguous electrocardiographic leads); or
- 3) Positive cardiac markers associated with non-ST segment myocardial infarction (NSTEMI) (CKMB or troponin I/T > local laboratory upper limit of normal values) within 24 hours after initial presentation
- This problem is borne out by the fact that 50% of cases identified in claims did not match registry cases … presumably in many cases because they did not meet the aforementioned criteria.
- Furthermore, we do not know whether data elements scraped from EHR, particularly vital signs, will perform as well as registry data. One can imagine that the first blood pressure or heart rate recorded might not be accurate, for example.
- The degree to which the distribution of observed and risk-adjusted differ (e.g. figure 4 vs. figure 5 in the technical report) is cause for concern given what is and isn't in the risk adjustment model.
- Missing data: 25% of registry cases could not be matched to Medicare claims and 50% of claims couldn't be matched to registry data. This is a problem.

2b2-3. Other Threats to Validity (Exclusions, Risk Adjustment):

- I don't have concerns about the risk adjustment development process.

- Risk adjustment appeared adequately developed and tested in 2014, but no current data presented for reassessment.
- Cerner EMR is excluded. Paper medical records are excluded. Non-Medicare patients are excluded. Type II MI is included. STEMI and NSTEMI are bunched together, yet the measure title suggests mortality for ALL patients...
- Social risk factors are not included. so while population-based, the bias of unique demographic population impacts on the measure is not accounted for.
- Please see previous comments re: inadequacy of risk adjustment. Key variables remain unmeasured to the extent that the validity of the measure cannot be assessed.
- Social risk factors not included in the model.

Criterion 3. Feasibility

Maintenance measures – no change in emphasis – implementation issues may be more prominent
3. Feasibility is the extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.
This is an hybrid measure. The developer indicates all data elements are in defined fields in a combination of electronic sources. The developer provided a feasibility scorecard; all components of the scorecard (i.e., data availability, data accuracy, data standards, and workflow) received a score of "3" (highest rating). The scorecard indicates that all critical data elements used for this measure (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) are currently feasible in existing EHR systems tested (Epic and Meditech).
The developer noted some initial time was required for a hospital to map the data elements in the measure specifications to their own EHR system. However, once these data elements were mapped, a hospital could submit many of these data elements for other hybrid measures, once implemented. Lastly, the developer indicated that there are no fees to use this measure.
Questions for the Committee:

- *Are the required data elements routinely generated and used during care delivery?*
- *Are the required data elements available in electronic form, e.g., EHR or other electronic sources?*
- *Is the data collection strategy ready to be put into operational use?*
- *If an eMeasure, does the eMeasure Feasibility Score Card demonstrate acceptable feasibility in multiple EHR systems and sites?*

Preliminary rating for feasibility: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee Pre-evaluation Comments: Criteria 3: Feasibility

3. Feasibility

- Feasibility is moderate because hospitals will need to initially map the data elements but this is a one-time requirement.
- This measure is not implemented. Though the developers said it is ready for implementation, it does not appear to be likely to be implemented anytime soon.
- As above, not feasible in current iteration.
- There is demonstrated feasibility for the use of claims-based data. The electronic medical record data capture was done with Epic and Meditech only so that the hundred of other EHRs were not subject to testing.
- For an EHR-based measure, the developer should demonstrate that it can be operationalized using actual EHR data. More important, I recommend deferring further development of this measure until s critical data elements such as whether a STEMI, whether true ACS vs. spurious circulating troponin, etc., can be reliably extracted from the medical record.

- Must map data elements in the measure to each hospital EHR system.

Criterion 4: Usability and Use

Maintenance measures – increased emphasis – much greater focus on measure use and usefulness, including both impact/improvement and unintended consequences

4a. Use (4a1. Accountability and Transparency; 4a2. Feedback on measure)

4a. Use evaluate the extent to which audiences (e.g., consumers, purchasers, providers, policymakers) use or could use performance results for both accountability and performance improvement activities.

4a.1. Accountability and Transparency. Performance results are used in at least one accountability application within three years after initial endorsement and are publicly reported within six years after initial endorsement (or the data on performance results are available). If not in use at the time of initial endorsement, then a credible plan for implementation within the specified timeframes is provided.

Current uses of the measure

Publicly reported? ☐ **Yes** ☒ **No**

Current use in an accountability program? ☐ **Yes** ☒ **No** ☐ **UNCLEAR**

OR

Planned use in an accountability program? ☒ **Yes** ☐ **No**

Accountability program details The developer indicates this measure is final and ready to be implemented by CMS in future regulations and is suitable for the HIQR program, the Hospital Value Based Payment (HVBP) program, or a future EPM under a CMMI program. However, the developer also noted that this measure was proposed and finalized into the Center for Medicare and Medicaid Services Innovation (CMMI) Advancing Care Coordination Through Episode Payment Models (EPM) five-year bundled payment model in January 2017 (82 FR 180). Yet, in December 2017, CMMI finalized the cancellation of the bundled payment model that included the hybrid AMI mortality measure (82 FR 57066). This measure was also signaled in the FY 2016 Hospital Inpatient Quality Reporting (HIQR) final rule in August 2015 (80 FR 49698).

4a.2. Feedback on the measure by those being measured or others. Three criteria demonstrate feedback: 1) those being measured have been given performance results or data, as well as assistance with interpreting the measure results and data; 2) those being measured and other users have been given an opportunity to provide feedback on the measure performance or implementation; 3) this feedback has been considered when changes are incorporated into the measure

Feedback on the measure by those being measured or others N/A

Additional Feedback: N/A

***Questions for the Committee**:*

- *How have (or can) the performance results be used to further the goal of high-quality, efficient healthcare?*
- *How has the measure been vetted in real-world settings by those being measured or others?*

Preliminary rating for Use: ☒ **Pass** ☐ **No Pass**

RATIONALE: This measure is ready to be implemented and was finalized into the Center for Medicare and Medicaid Services Innovation (CMMI) Advancing Care Coordination Through Episode Payment Models (EPM) five-year bundled payment model in January 2017 (82 FR 180). Yet, in December 2017, CMMI finalized the cancellation of the bundled payment model.

4b. Usability (4a1. Improvement; 4a2. Benefits of measure)

4b. Usability evaluate the extent to which audiences (e.g., consumers, purchasers, providers, policymakers) use or could use performance results for both accountability and performance improvement activities.
4b.1 Improvement. Progress toward achieving the goal of high-quality, efficient healthcare for individuals or populations is demonstrated.
Improvement results N/A. Measure is currently not implemented
4b2. Benefits vs. harms. Benefits of the performance measure in facilitating progress toward achieving high-quality, efficient healthcare for individuals or populations outweigh evidence of unintended negative consequences to individuals or populations (if such evidence exists).
Unexpected findings (positive or negative) during implementation N/A. Measure is currently not implemented
Potential harms The developer indicates that they did not identify any unintended consequences during measure development, model testing, or testing the risk variables in hospital settings.
Additional Feedback:

***Questions for the Committee**:*

- *Do the benefits of the measure outweigh any potential unintended consequences?*

Preliminary rating for Usability and use: ☐ **High** ☐ **Moderate** ☐ **Low** ☒ **Insufficient**

RATIONALE: No improvement results were provided or unexpected findings because this measure is not yet implemented. This was finalized into the Center for Medicare and Medicaid Services Innovation (CMMI) Advancing Care Coordination Through Episode Payment Models (EPM) five-year bundled payment model in January 2017 (82 FR 180). Yet, in December 2017, CMMI finalized the cancellation of the bundled payment model.

Committee Pre-evaluation Comments: Criteria 4: Usability and Use

4a. Use

- The measure is ready for use. It has not been used yet.
- Not currently implemented.
- Should to be used for payment due to aforementioned considerations
- The measure in its claims based has been used so that it is presumed that with the EHR data it will continue to be used; and meet the accountability and transparency requirements.
- Not currently in use for public reporting but planned.

4b. Usability

- Usability is not known because it hasn't been implemented.
- Questionable usability due to unaddressed issues with EMR data gathering and reconciliation with claims data.
- Could result in significant negative effects given its half-baked nature at this time. It will likely work well in the future, as hybrid e-measures are clearly the way to improve this field, it is not ready for prime time now.
- The concept of reducing mortality from all causes requires an accountability structure not specified or mentioned in the measure. will that change in structure of health care place greater resources on care management for the first 30 days after a CABG or PTCA, rather than look for the long-island approach.
- As defined, the measure might encourage up-coding to an AMI diagnosis among institutions not already doing so.
- Not currently implemented. Perhaps needs further use to demonstrate usability.

Criterion 5: Related and Competing Measures

Related or competing measures

- This measure is related to:
    - 0230: Hospital 30-day, all-cause, risk-standardized mortality rate (RSMR) following acute myocardial infarction (AMI) hospitalization for patients 18 and older [NOTE: This is the claims-based measure of this submission]
- This measure is competing with:
    - 0730: Acute Myocardial Infarction (AMI) Mortality Rate.

Harmonization

The developer notes that the measure specifications are, by design, not completely harmonized in that the current measure uses clinical data elements collected from EHR for risk adjustment, and the measures listed above use claims data for risk adjustment. Additionally, the outcome in measure #0730 is inpatient mortality rather than 30-day mortality. Inpatient mortality rates can be influenced by hospital length of stay, so 30-day measures that establish a standard follow-up period are more appropriate for profiling a diverse group of hospitals. The measures listed above have target populations aged 18+, whereas the current measure's target population is age 65+. The exclusion criteria of the current measure are largely similar to those of measure #0230.

Committee Pre-evaluation Comments: Criterion 5: Related and Competing Measures

- None

Public and Member Comments

Comments and Member Support/Non-Support Submitted as of: June 12, 2018

No comments have been submitted as of this date.

Developer Submission

Measure Information

This document contains the information submitted by measure developers/stewards, but is organized according to NQF's measure evaluation criteria and process. The item numbers refer to those in the submission form but may be in a slightly different order here. In general, the item numbers also reference the related criteria (e.g., item 1b.1 relates to sub criterion 1b).

Brief Measure Information

NQF #: 2473e

Corresponding Measures:

De.2. Measure Title: Hybrid hospital 30-day, all-cause, risk-standardized mortality rate (RSMR) following acute myocardial infarction (AMI)

Co.1.1. Measure Steward: Centers for Medicare & Medicaid Services (CMS)

De.3. Brief Description of Measure: This measure estimates a hospital-level 30-day, all-cause, risk-standardized mortality rate (RSMR) for patients discharged from the hospital with a principal discharge diagnosis of acute myocardial infarction (AMI). The outcome is all-cause 30-day mortality, defined as death from any cause within 30 days of the index admission date, including in-hospital death, for AMI patients. The target population is Medicare Fee-for-Service beneficiaries who are 65 years or older.

This Hybrid AMI mortality measure was developed de novo. This measure is harmonized with the Centers for Medicare and Medicaid Services' (CMS's) current publicly reported claims-only measure, hospital 30-day, all-cause, risk-standardized mortality rate (RSMR) following acute myocardial infarction (AMI) (NQF #2473). The measure is referred to as a hybrid because it is CMS's intention to calculate the measure using two data sources: Medicare fee-for-service (FFS) administrative claims and clinical electronic health record (EHR) data.

1b.1. Developer Rationale: The goal of this measure is to improve patient outcomes by providing patients, physicians, and hospitals with information about hospital-level, risk-standardized mortality rates following hospitalization for AMI. Measurement of patient outcomes allows for a broad view of quality of care that encompasses more than what can be captured by individual process-of-care measures. Complex and critical aspects of care, such as communication between providers, prevention of and response to complications, patient safety, and coordinated transitions to the outpatient environment, all contribute to patient outcomes but are difficult to measure by individual process measures. The goal of outcomes measurement is to risk-adjust for patients' conditions at the time of hospital admission and then evaluate patient outcomes. This measure was developed to identify institutions whose performance is better or worse than what would be expected based on their patient case mix, and therefore promote hospital quality improvement and better inform consumers about care quality.

Additionally, AMI mortality is a priority area for outcomes measure development as it is a costly and common condition. Hospital mortality is an outcome that is likely attributable to care processes and is an important outcome for patients. Measuring and reporting mortality rates will inform health care providers about opportunities to improve care, strengthen incentives for quality improvement, and ultimately improve the quality of care received by Medicare patients. The measure will also provide patients with information that could guide their choices. Furthermore, the measure will increase transparency for consumers and has the potential to lower health care costs associated with mortality.

This Hybrid AMI mortality measure incorporates clinical data elements pulled from the EHR in risk adjustment of the mortality model. Some benefits of including the clinical data elements are:

1. Inclusion of patient-level clinical data related to severity of illness is responsive to providers who continue to express preference for using patient-level clinical data, and provides an opportunity to incorporate clinical data into outcome measures.

2. Hospitals will increasingly use EHR data to assess severity of illness and patients' risk of poor outcomes. This provides an opportunity to align the measure with clinical decision support systems that many providers utilize to alert care teams about patients at increased risk of poor outcomes in real time during the inpatient stay.

3. Collecting a simple core set of clinical data elements that perform well as risk-adjustment variables (for illness severity) across conditions can greatly reduce the cost and effort of future measure development, improve harmonization, and create opportunity for longitudinal assessment of patient status and quality of care across settings.

4. These core clinical data elements will provide measure developers with a standard set of reliable data that can be used as a starting place when building risk-adjustment models for quality measures using clinical data.

S.4. Numerator Statement: The outcome is all-cause 30-day mortality, defined as death from any cause within 30 days of the index admission date, including in-hospital death, for patients with a principal discharge diagnosis of AMI.

S.6. Denominator Statement: The cohort includes inpatient admissions for Medicare FFS patients 65 years and older who were discharged from non-federal, short-term, acute care hospitals with a principal discharge diagnosis of AMI.

Additional details are provided in S.7 Denominator Details.

S.8. Denominator Exclusions: The mortality measure excludes index hospitalizations that meet any of the following exclusion criteria:

1. Discharged alive on the day of admission or the following day, who were not transferred to another acute care facility;
2. With inconsistent or unknown vital status or other unreliable demographic (age and gender) data;
3. Enrolled in the Medicare hospice program any time in the 12 months prior to the index admission, including the first day of the index admission; or

4. Discharged against medical advice (AMA).

After exclusions #1-4 are applied, the measure randomly selects one index admission per patient per year for inclusion in the cohort so that each episode of care is mutually independent with the same probability of the outcome. Additional admissions within that year are excluded. For each patient, the probability of death increases with each subsequent admission and therefore the episodes of care are not mutually independent. For the three-year combined data, when index admissions occur during the transition between measure reporting periods (June and July of each year) and both are randomly selected for inclusion in the measure, the measure includes only the June admission. July admissions are excluded to avoid assigning a single death to two admissions.

De.1. Measure Type: Outcome

S.17. Data Source: Claims, Electronic Health Data, Electronic Health Records, Other, Registry Data

S.20. Level of Analysis: Facility

IF Endorsement Maintenance – Original Endorsement Date: Sep 08, 2014 **Most Recent Endorsement Date:** Sep 08, 2014

IF this measure is included in a composite, NQF Composite#/title:

IF this measure is paired/grouped, NQF#/title:

De.4. IF PAIRED/GROUPED, what is the reason this measure must be reported with other measures to appropriately interpret results? N/A

1. Evidence and Performance Gap – Importance to Measure and Report

Extent to which the specific measure focus is evidence-based, important to making significant gains in healthcare quality, and improving health outcomes for a specific high-priority (high-impact) aspect of healthcare where there is variation in or overall less-than-optimal performance. ***Measures must be judged to meet all sub criteria to pass this criterion and be evaluated against the remaining criteria.***

1a. Evidence to Support the Measure Focus – See attached Evidence Submission Form

Del18eHOY4HybridAMIMortalityEndorsementMaintenanceEvidenceAttachment03262018.docx

1a.1 For Maintenance of Endorsement: Is there new evidence about the measure since the last update/submission? Do not remove any existing information. If there have been any changes to evidence, the Committee will consider the new evidence. Please use the most current version of the evidence attachment (v7.1). Please use red font to indicate updated evidence.

No

1a Evidence (subcriterion 1a)

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

Measure Title: Hybrid Hospital 30-Day, All-Cause, Risk-Standardized Mortality Rate (RSMR) following Acute Myocardial Infarction (AMI)

IF the measure is a component in a composite performance measure, provide the title of the Composite Measure here: n/a

Date of Submission: 4/9/2018

Instructions

- *Complete 1a.1 and 1a.2 for all measures. If instrument-based measure, complete 1a.3.*
- *Complete **EITHER 1a.2, 1a.3 or 1a.4** as applicable for the type of measure and evidence.*
- *For composite performance measures:*
  - *A separate evidence form is required for each component measure unless several components were studied together.*
  - *If a component measure is submitted as an individual performance measure, attach the evidence form to the individual measure submission.*
- All information needed to demonstrate meeting the evidence subcriterion (1a) must be in this form. An appendix of *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.
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.

Note: The information provided in this form is intended to aid the Standing Committee and other stakeholders in understanding to what degree the evidence for this measure meets NQF's evaluation criteria.

1a. Evidence to Support the Measure Focus

The measure focus is evidence-based, demonstrated as follows:

- Outcome: [3] Empirical data demonstrate a relationship between the outcome and at least one healthcare structure, process, intervention, or service. If not available, wide variation in performance can be used as evidence, assuming the data are from a robust number of providers and results are not subject to systematic bias.
- Intermediate clinical outcome: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured intermediate clinical outcome leads to a desired health outcome.
- Process: [5] a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured process leads to a desired health outcome.

- Structure: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured structure leads to a desired health outcome.
- Efficiency: [6] evidence not required for the resource use component.
- For measures derived from patient reports, evidence should demonstrate that the target population values the measured outcome, process, or structure and finds it meaningful.
- Process measures incorporating Appropriate Use Criteria: See NQF's guidance for evidence for measures, in general; guidance for measures specifically based on clinical practice guidelines apply as well.

Notes

3. Generally, rare event outcomes do not provide adequate information for improvement or discrimination; however, serious reportable events that are compared to zero are appropriate outcomes for public reporting and quality improvement.

4. The preferred systems for grading the evidence are the Grading of Recommendations, Assessment, Development and Evaluation (GRADE) guidelines and/or modified GRADE.

5. Clinical care processes typically include multiple steps: assess → identify problem/potential problem → choose/plan intervention (with patient input) → provide intervention → evaluate impact on health status. If the measure focus is one step in such a multistep process, the step with the strongest evidence for the link to the desired outcome should be selected as the focus of measurement. Note: A measure focused only on collecting PROM data is not a PRO-PM.

6. Measures of efficiency combine the concepts of resource use and quality (see NQF's Measurement Framework: Evaluating Efficiency Across Episodes of Care; AQA Principles of Efficiency Measures).

1a.1.This is a measure of: (*should be consistent with type of measure entered in De.1*)

Outcome

☒Outcome: Hospital 30-day mortality

☐Patient-reported outcome (PRO):

PROs include HRQoL/functional status, symptom/symptom burden, experience with care, health-related behaviors. (*A PRO-based performance measure is not a survey instrument. Data may be collected using a survey instrument to construct a PRO measure.)*

☐ Intermediate clinical outcome (*e.g., lab value*):

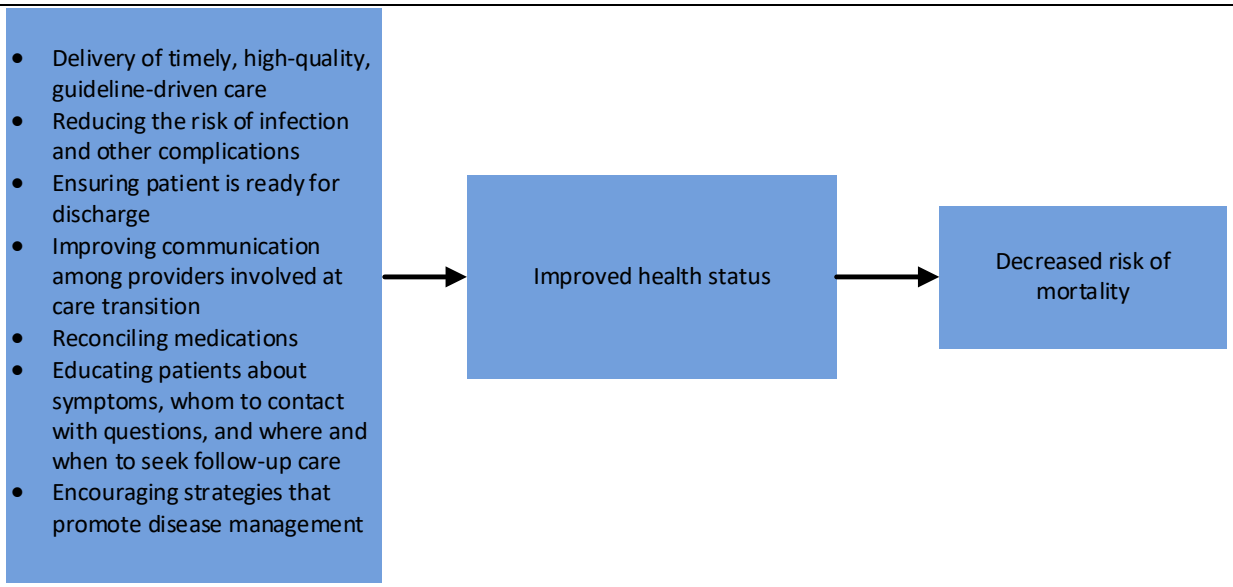
☐Process:

☐ Appropriate use measure:

☐Structure:

☐Composite:

1a.2 LOGIC MODEL Diagram or briefly describe the steps between the healthcare structures and processes (e.g., interventions, or services) and the patient's health outcome(s). The relationships in the diagram should be easily understood by general, non-technical audiences. Indicate the structure, process or outcome being measured.



The goal of this measure is to improve patient outcomes by providing patients, physicians, and hospitals with information about hospital-level, risk-standardized mortality rates following hospitalization for acute myocardial infarction (AMI). Measurement of patient outcomes allows for a broad view of quality of care that encompasses more than what can be captured by individual process-of-care measures. Complex and critical aspects of care, such as communication between providers, prevention of, and response to, complications, patient safety and coordinated transitions to the outpatient environment, all contribute to patient outcomes but are difficult to measure by individual process measures. The goal of outcomes measurement is to risk-adjust for patients' conditions at the time of hospital admission and then evaluate patient outcomes. This mortality measure was developed to identify institutions, whose performance is better or worse than would be expected based on their patient case-mix, and therefore promote hospital quality improvement and better inform consumers about care quality.

Because it is developed for use in EHRs, this measure can utilize detailed clinical data without requiring the investment of resources currently needed to collect registry or medical record-abstracted data. This Hybrid measure of AMI mortality responds to stakeholders' interest in using clinical data from medical records for risk adjustment in outcome measures. This measure will provide critical insight into AMI outcomes across hospitals, using clinical data for risk adjustment without undue burden on hospitals

1a.3 Value and Meaningfulness: IF this measure is derived from patient report, provide evidence that the target population values the measured ***outcome, process, or structure*** and finds it meaningful. (Describe how and from whom their input was obtained.)

****RESPOND TO ONLY ONE SECTION BELOW -EITHER 1a.2, 1a.3 or 1a.4) ****

1a.2 FOR OUTCOME MEASURES including PATIENT REPORTED OUTCOMES - Provide empirical data demonstrating the relationship between the outcome (or PRO) to at least one healthcare structure, process, intervention, or service.

The measures assess mortality within a 30-day period from the date of the index admission. From a patient perspective, death is a critical outcome regardless of cause. The measures use a 30-day time frame because older adult patients are more vulnerable to adverse health outcomes occurring during this time. Death within 30 days of the start of the admission can be influenced by hospital care and the early transition to the non-acute care setting. The 30-day time frame is a clinically meaningful period for hospitals to collaborate with their communities in an effort to reduce mortality.

Complex and critical aspects of care – such as communication between providers, prevention of and response to complications, patient safety, and coordinated transitions to the outpatient environment – all contribute to patient outcomes but are difficult to measure by individual process measures. Furthermore, recent work has identified specific strategies utilized by hospitals that achieve low AMI mortality rates (Bradley et al., 2012; Curry et al., 2011). These strategies include having cardiologists on site, the presence of physician and nurse champions, and promoting strong

communication and coordination across disciplines and departments. This work demonstrates the relationship between hospital organizational factors and performance on the AMI mortality measures and supports the ability of hospitals to impact these rates.

Reference:

1. Bradley EH, Curry LA, Spatz ES, Herrin J, Cherlin EJ, Curtis JP, Thompson JW, Ting HH, Wang Y, Krumholz HM. Hospital strategies for reducing risk-standardized mortality rates in acute myocardial infarction. Ann Intern Med. 2012 May 1;156(9):618-26.

2. Dharmarajan K, Hsieh AF, Kulkarni VT, et al. Trajectories of risk after hospitalization for heart failure, acute myocardial infarction, or pneumonia: retrospective cohort study. BMJ (Clinical research ed). 2015;350:h411.

3. Curry LA, Spatz E, Cherlin E, Thompson JW, Berg D, Ting HH, Decker C, Krumholz HM, Bradley EH. What distinguishes top-performing hospitals in acute myocardial infarction mortality rates? A qualitative study. Ann Intern Med. 2011 Mar 15;154(6):384-90.

4. Drye E, Normand S, Wang Y, et al. Comparison of hospital risk-standardized mortality rates calculated by using in-hospital and 30-day models: an observational study with implications for hospital profiling. Annals of internal medicine. 2012;156(1 Pt 1):19-26.

1a.3. SYSTEMATIC REVIEW(SR) OF THE EVIDENCE (for INTERMEDIATE OUTCOME, PROCESS, OR STRUCTURE PERFORMANCE MEASURES, INCLUDING THOSE THAT ARE INSTRUMENT-BASED) If the evidence is not based on a systematic review go to section 1a.4) If you wish to include more than one systematic review, add additional tables.

What is the source of the systematic review of the body of evidence that supports the performance measure? A systematic review is a scientific investigation that focuses on a specific question and uses explicit, prespecified scientific methods to identify, select, assess, and summarize the findings of similar but separate studies. It may include a quantitative synthesis (meta-analysis), depending on the available data. (IOM)

☐Clinical Practice Guideline recommendation (with evidence review)

☐US Preventive Services Task Force Recommendation

☐Other systematic review and grading of the body of evidence (*e.g., Cochrane Collaboration, AHRQ Evidence Practice Center*)

☐Other

Source of Systematic Review: • Title • Author • Date • Citation, including page number • URL	
Quote the guideline or recommendation verbatim about the process, structure or intermediate outcome being measured. If not a guideline, summarize the conclusions from the SR.	
Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	
Provide all other grades and definitions from the evidence grading system	
Grade assigned to the **recommendation** with definition of the grade	

Provide all other grades and definitions from the recommendation grading system	
Body of evidence: • Quantity – how many studies? • Quality – what type of studies?	
Estimates of benefit and consistency across studies	
What harms were identified?	
Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	

1a.4 OTHER SOURCE OF EVIDENCE

If source of evidence is NOT from a clinical practice guideline, USPSTF, or systematic review, please describe the evidence on which you are basing the performance measure.

1a.4.1 Briefly SYNTHESIZE the evidence that supports the measure. A list of references without a summary is not acceptable.

1a.4.2 What process was used to identify the evidence?

1a.4.3. Provide the citation(s) for the evidence.

1b. Performance Gap

Demonstration of quality problems and opportunity for improvement, i.e., data demonstrating:

- considerable variation, or overall less-than-optimal performance, in the quality of care across providers; and/or
- Disparities in care across population groups.

1b.1. Briefly explain the rationale for this measure *(e.g., how the measure will improve the quality of care, the benefits or improvements in quality envisioned by use of this measure)*

If a COMPOSITE (e.g., combination of component measure scores, all-or-none, any-or-none), SKIP this question and answer the composite questions.

The goal of this measure is to improve patient outcomes by providing patients, physicians, and hospitals with information about hospital-level, risk-standardized mortality rates following hospitalization for AMI. Measurement of patient outcomes allows for a broad view of quality of care that encompasses more than what can be captured by individual process-of-care measures. Complex and critical aspects of care, such as communication between providers, prevention of and response to complications, patient safety, and coordinated transitions to the outpatient environment, all contribute to patient outcomes but are difficult to measure by individual process measures. The goal of outcomes measurement is to risk-adjust for patients' conditions at the time of hospital admission and then evaluate patient outcomes. This measure was developed to identify institutions whose performance is better or worse than what would be expected based on their patient case mix, and therefore promote hospital quality improvement and better inform consumers about care quality.

Additionally, AMI mortality is a priority area for outcomes measure development as it is a costly and common condition. Hospital mortality is an outcome that is likely attributable to care processes and is an important outcome for patients. Measuring and reporting mortality rates will inform health care providers about opportunities to improve care, strengthen incentives for quality improvement, and ultimately improve the quality of care received by Medicare patients. The measure will also provide patients with information that could guide their choices. Furthermore, the measure will increase transparency for consumers and has the potential to lower health care costs associated with mortality.

This Hybrid AMI mortality measure incorporates clinical data elements pulled from the EHR in risk adjustment of the mortality model. Some benefits of including the clinical data elements are:

1. Inclusion of patient-level clinical data related to severity of illness is responsive to providers who continue to express preference for using patient-level clinical data, and provides an opportunity to incorporate clinical data into outcome measures.

2. Hospitals will increasingly use EHR data to assess severity of illness and patients' risk of poor outcomes. This provides an opportunity to align the measure with clinical decision support systems that many providers utilize to alert care teams about patients at increased risk of poor outcomes in real time during the inpatient stay.

3. Collecting a simple core set of clinical data elements that perform well as risk-adjustment variables (for illness severity) across conditions can greatly reduce the cost and effort of future measure development, improve harmonization, and create opportunity for longitudinal assessment of patient status and quality of care across settings.

4. These core clinical data elements will provide measure developers with a standard set of reliable data that can be used as a starting place when building risk-adjustment models for quality measures using clinical data.

1b.2. Provide performance scores on the measure as specified (current and over time) at the specified level of analysis. *(This is required for maintenance of endorsement. Include mean, std dev, min, max, interquartile range, scores by decile. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities include.) This information also will be used to address the sub-criterion on improvement (4b1) under Usability and Use.*

We analyzed variation in RSMRs among the hospitals in the development dataset – i.e., hospitals participating in the ACTION Registry(R)–GWTG(TM) (AR-G), for clinical data, merged with CMS Medicare claims and enrollment data – for the 30-day mortality outcome.

The development cohort includes AMI discharges for patients aged 65 and older from January 1 - December 31, 2009 who were discharged from hospitals participating in the AR-G and who were enrolled in Medicare. It includes 20,540 admissions from 280 hospitals. AMI RSMRs vary among hospitals, with a mean of 10.8%, a standard deviation of 0.006, and a range of 9.6% to 13.1%. The interquartile range is 10.3% to 11.1%. The set of hospitals included is likely to have a narrow range of performance due to their participation in the AR-G registry. The mean score by decile is as follows:

Decile of RSMR	Mean RSMR
1	0.100
2	0.103
3	0.105
4	0.107
5	0.107
6	0.108
7	0.109
8	0.110
9	0.112
10	0.118

1b.3. If no or limited performance data on the measure as specified is reported in 1b2, then provide a summary of data from the literature that indicates opportunity for improvement or overall less than optimal performance on the specific focus of measurement.

CMS currently publicly reports a claims-based measure of AMI mortality. The results for this measure, as reported in the 2017 update to the Hospital Compare website, are based on RSMRs calculated for AMI admissions among Medicare fee-for-service patients aged 65 and older from July 2013 – June 2016. It includes 487,646 admissions from 4,310 hospitals. For the most recently reported three years of data (July 2013 - June 2016), the mean hospital RSMR was 13.7%, with a range of 9.7% to 18.0%. The interquartile range was 13.2% to 14.1%.

Trends indicate a decrease in both the Publicly Reported National Rate and Median Hospital Rate using the claims-based measure of AMI mortality.

Year	Publicly Reported National Rate	Median Hospital Rate	Range
2012	---	15.7%	10.0-21.5
2013	15.2%	15.1%	9.4-21.0
2014	14.9%	14.8%	9.4-20.2
2015	14.2%	14.1%	9.9-15.7
2016	14.1%	14.0%	9.4-20.0
2017	13.6%	13.5%	9.7-18.0

Furthermore, recent work has identified specific strategies utilized by hospitals that achieve low AMI mortality rates (Bradley et al., 2012; Curry et al., 2011). This work demonstrates the relationship between hospital organizational factors and performance on the AMI mortality measures and supports the ability of hospitals to impact these rates.

References:

Bradley EH, Curry LA, Spatz ES, Herrin J, Cherlin EJ, Curtis JP, Thompson JW, Ting HH, Wang Y, Krumholz HM. Hospital strategies for reducing risk-standardized mortality rates in acute myocardial infarction. Ann Intern Med. 2012 May 1;156(9):618-26.

Curry LA, Spatz E, Cherlin E, Thompson JW, Berg D, Ting HH, Decker C, Krumholz HM, Bradley EH. What distinguishes top-performing hospitals in acute myocardial infarction mortality rates? A qualitative study. Ann Intern Med. 2011 Mar 15;154(6):384-90.

1b.4. Provide disparities data from the measure as specified (current and over time) by population group, e.g., by race/ethnicity, gender, age, insurance status, socioeconomic status, and/or disability. *(This is required for maintenance of endorsement. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included.) For measures that show high levels of performance, i.e., "topped out", disparities data may demonstrate an opportunity for improvement/gap in care for certain sub-populations. This information also will be used to address the sub-criterion on improvement (4b1) under Usability and Use.*

We analyzed whether disparities in performance on this measure exist at the hospital level.

To identify potential disparities related to race, we examined the relationship between RSMR and hospital proportion of African-American patients among all hospitals included in the merged AR-G-CMS dataset used for measure development. We used the 2009 Medicare Provider Analysis and Review (MEDPAR) file to calculate the percentage of African-American patients treated at each hospital, using all patients admitted to each hospital. We classified hospitals into quintiles based on their proportion of African-American patients, with the lowest and highest quintile consisting of hospitals with lowest and highest proportions of African-American patients, respectively.

Analyses demonstrated that median RSMRs and the distributions of RSMRs were consistent across quintiles. Specifically, the median RSMR for hospitals in the lowest quintile was 10.8%, and the median RSMR for hospitals in the highest quintile was 10.8%. This analysis suggests that many hospitals with a high proportion of African-American patients can and do perform well on the measure.

To identify potential disparities related to socioeconomic status (SES), we examined the relationship between RSMR and hospital proportion of dual eligible patients. We used the 2009 MEDPAR file to calculate the percentage of dual eligible patients treated at each hospital. We used Medicaid eligibility status identified in the Medicare Enrollment Database as a proxy for SES. This approach is consistent with prior research as well as National Quality Forum (NQF) recommendations (http://www.qualityforum.org/Publications/2011/07/National_Voluntary_Consensus_Standards_for_Patient_Outcomes_2009.aspx). Hospitals were categorized into quintiles based on their proportion of dual eligible patients, with the lowest and highest quintile consisting of hospitals with lowest and highest proportions of dual eligible patients, respectively. Analyses showed that median RSMRs were consistent across quintiles of hospitals based on the hospital proportion of dual eligible patients. Specifically, the median RSMR for hospitals in the lowest quintile was 10.8%, and the median RSMR for hospitals in the highest quintile was 10.9%. The distributions were also consistent across quintiles. These results indicate that hospitals with high proportions of dual eligible patients can and do perform as well on the measure as hospitals with lower proportions of dual eligible patients.

1b.5. If no or limited data on disparities from the measure as specified is reported in 1b.4, then provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement. Include citations. Not necessary if performance data provided in 1b.4

N/A

2. Reliability and Validity—Scientific Acceptability of Measure Properties

Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. ***Measures must be judged to meet the sub criteria for both reliability and validity to pass this criterion and be evaluated against the remaining criteria.***

2a.1. Specifications The measure is well defined and precisely specified so it can be implemented consistently within and across organizations and allows for comparability. eMeasures should be specified in the Health Quality Measures Format (HQMF) and the Quality Data Model (QDM).

De.5. Subject/Topic Area *(check all the areas that apply):*

Cardiovascular : Coronary Artery Disease (AMI)

De.6. Non-Condition Specific*(check all the areas that apply):*

Care Coordination, Safety, Safety : Complications

De.7. Target Population Category *(Check all the populations for which the measure is specified and tested if any):*

Elderly, Populations at Risk

S.1. Measure-specific Web Page *(Provide a URL link to a web page specific for this measure that contains current detailed specifications including code lists, risk model details, and supplemental materials. Do not enter a URL linking to a home page or to general information.)*

https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/HospitalQualityInits/Downloads/Core-Clinical-Data-Elements-and-Hybrid-Measures.zip

S.2a. If this is an eMeasure, HQMF specifications must be attached. Attach the zipped output from the eMeasure authoring tool (MAT) - if the MAT was not used, contact staff. (Use the specification fields in this online form for the plain-language description of the specifications)

This is an eMeasure **Attachment:** CCDE_AMI_Mortality_2016_Final_Specifications-636510023794915089.zip

S.2b. Data Dictionary, Code Table, or Value Sets *(and risk model codes and coefficients when applicable) must be attached. (Excel or csv file in the suggested format preferred - if not, contact staff)*

Attachment **Attachment:** HOY4_Hybrid_AMI_Mortality_Data_Dictionary_v1.0.xls

S.2c. Is this an instrument-based measure (i.e., data collected via instruments, surveys, tools, questionnaires, scales, etc.)? Attach copy of instrument if available.

No, this is not an instrument-based measure **Attachment:**

S.2d. Is this an instrument-based measure (i.e., data collected via instruments, surveys, tools, questionnaires, scales, etc.)? Attach copy of instrument if available.

Not an instrument-based measure

S.3.1. For maintenance of endorsement: Are there changes to the specifications since the last updates/submission. If yes, update the specifications for S1-2 and S4-22 and explain reasons for the changes in S3.2.

No

S.3.2. For maintenance of endorsement, please briefly describe any important changes to the measure specifications since last measure update and explain the reasons.

N/A

S.4. Numerator Statement *(Brief, narrative description of the measure focus or what is being measured about the target population, i.e., cases from the target population with the target process, condition, event, or outcome) DO NOT include the rationale for the measure.*

IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm (S.14).

The outcome is all-cause 30-day mortality, defined as death from any cause within 30 days of the index admission date, including in-hospital death, for patients with a principal discharge diagnosis of AMI.

S.5. Numerator Details *(All information required to identify and calculate the cases from the target population with the target process, condition, event, or outcome such as definitions, time period for data collection, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*

IF an OUTCOME MEASURE, describe how the observed outcome is identified/counted. Calculation of the risk-adjusted outcome should be described in the calculation algorithm (S.14).

The measure outcome is death from any cause within 30 days of the admission date of the index admission. As currently specified, we identify deaths for Medicare FFS patients 65 years and older in the Medicare Enrollment Database (EDB).

S.6. Denominator Statement *(Brief, narrative description of the target population being measured)*

The cohort includes inpatient admissions for Medicare FFS patients 65 years and older who were discharged from non-federal, short-term, acute care hospitals with a principal discharge diagnosis of AMI.

Additional details are provided in S.7 Denominator Details.

S.7. Denominator Details *(All information required to identify and calculate the target population/denominator such as definitions, time period for data collection, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b.)*

IF an OUTCOME MEASURE, describe how the target population is identified. Calculation of the risk-adjusted outcome should be described in the calculation algorithm (S.14).

To be included in the measure cohort, patients must meet the following inclusion criteria:

1. Had a principal discharge diagnosis of AMI;
2. Enrolled in Medicare FFS Part A and Part B for the first 12 months prior to the date of admission, and enrolled in Part A during the index admission;
3. Aged 65 or over; and
4. Not transferred from another acute care facility.

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

S.8. Denominator Exclusions *(Brief narrative description of exclusions from the target population)*

The mortality measure excludes index hospitalizations that meet any of the following exclusion criteria:

1. Discharged alive on the day of admission or the following day, who were not transferred to another acute care facility;
2. With inconsistent or unknown vital status or other unreliable demographic (age and gender) data;
3. Enrolled in the Medicare hospice program any time in the 12 months prior to the index admission, including the first day of the index admission; or
4. Discharged against medical advice (AMA).

After exclusions #1-4 are applied, the measure randomly selects one index admission per patient per year for inclusion in the cohort so that each episode of care is mutually independent with the same probability of the outcome. Additional admissions within that year are excluded. For each patient, the probability of death increases with each subsequent admission and therefore the episodes of care are not mutually independent. For the three-year combined data, when index admissions occur during the transition between measure reporting periods (June and July of each year) and both are randomly selected for inclusion in the measure, the measure includes only the June admission. July admissions are excluded to avoid assigning a single death to two admissions.

S.9. Denominator Exclusion Details *(All information required to identify and calculate exclusions from the denominator such as definitions, time period for data collection, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b.)*

1. Discharged alive on the day of admission or the following day who were not transferred to another acute care facility

Rationale: It is unlikely that these patients had clinically significant AMI. This is determined from the claim

2. Inconsistent or unknown vital status or other unreliable demographic data

Rationale: We do not include stays for patients where the age is greater than 115 (indicated in the claim), where the gender is neither male nor female (indicated in the claim), where the admission date in the claim is after the date of death in the Medicare Enrollment Database, or where the date of death occurs before the date of discharge but the patient was discharged alive as indicated in the claim.

3. Enrolled in the Medicare hospice program any time in the 12 months prior to the index admission, including the first day of the index admission

Rationale: These patients are likely continuing to seek comfort measures only, so mortality is not necessarily an adverse outcome or signal of poor quality care. This is indicated in the claim.

4. Discharged against medical advice

Rationale: Providers did not have the opportunity to deliver full care and prepare the patient for discharge. This is determined from the discharge disposition in the claim.

S.10. Stratification Information *(Provide all information required to stratify the measure results, if necessary, including the stratification variables, definitions, specific data collection items/responses, code/value sets, and the risk-model covariates and coefficients for the clinically-adjusted version of the measure when appropriate – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format with at S.2b.)*

N/A

S.11. Risk Adjustment Type (Select type. Provide specifications for risk stratification in measure testing attachment)

Statistical risk model

If other:

S.12. Type of score:

Rate/proportion

If other:

S.13. Interpretation of Score *(Classifies interpretation of score according to whether better quality is associated with a higher score, a lower score, a score falling within a defined interval, or a passing score)*

Better quality = Lower score

S.14. Calculation Algorithm/Measure Logic (*Diagram or describe the calculation of the measure score as an ordered sequence of steps including identifying the target population; exclusions; cases meeting the target process, condition, event, or outcome; time period for data, aggregating data; risk adjustment; etc.)*

The measure estimates hospital-level 30-day all-cause RSMRs following AMI using hierarchical logistic regression models. In brief, the approach simultaneously models data at the patient and hospital levels to account for variance in patient outcomes within and between hospitals (Normand and Shahian, 2007). At the patient level, it models the log-odds of mortality within 30 days of discharge using age, sex, selected clinical covariates, and a hospital-specific intercept. At the hospital level, the approach models the hospital-specific effects as arising from a normal distribution. The hospital effect represents the underlying risk of a readmission at the hospital, after accounting for patient risk. The hospital-specific effects are given a distribution 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 effects should be identical across all hospitals.

This measure uses risk variables from electronic health records (EHR). The model adjusts for case-mix differences based on the clinical status of patients at the time of admission. Clinical risk-adjustment variables are the first values collected during the inpatient episode of care, including values collected in the emergency department or outpatient department in the 24 hours prior to inpatient admission.

Risk adjustment variables:

Age (years, continuous) for patients aged 65 or over

Heart rate

Systolic blood pressure

Creatinine

Troponin level

The RSMR is calculated as the ratio of the number of “predicted” to the number of “expected” deaths, multiplied by the national unadjusted mortality rate. For each hospital, the numerator of the ratio (“predicted”) is the number of deaths within 30 days predicted on the basis of the hospital’s performance with its observed case mix, and the denominator (“expected”) is the number of deaths expected on the basis of the nation’s performance with that hospital’s case mix. This approach is analogous to a ratio of “observed” to “expected” used in other types of statistical analyses. It conceptually allows for a comparison of a particular hospital’s performance given its case mix to an average hospital’s performance with the same case mix. Thus, a lower ratio indicates lower-than-expected mortality or better quality and a higher ratio indicates higher-than-expected mortality or worse quality.

The “predicted” number of deaths (the numerator) is calculated by using the coefficients estimated by regressing the risk factors and the hospital-specific intercept on the risk of mortality. The estimated hospital-specific intercept is added to the sum of the estimated regression coefficients multiplied by the patient characteristics. The results are transformed and summed over all patients attributed to a hospital to get a predicted value. The “expected” number of deaths (the denominator) is obtained in the same manner, but a common intercept using all hospitals in the sample is added in place of the hospital-specific intercept. The results are transformed and summed over all patients in the hospital to get an expected value. To assess hospital performance for each reporting period, we re-estimate the model coefficients using the years of data in that period.

This calculation transforms the ratio of predicted over expected into a rate that is compared to the national observed readmission rate. The hierarchical logistic regression models are described fully in the original methodology report for the claims-only AMI mortality measure (Krumholz et al., 2005).

Reference:

1. Normand S-LT, Shahian DM. 2007. Statistical and Clinical Aspects of Hospital Outcomes Profiling. Stat Sci 22(2): 206-226.

2. Krumholz H, Normand S, Galusha D, et al. Risk-Adjustment Models for AMI and HF 30-Day Mortality Methodology. 2005.

S.15. Sampling *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*

IF an instrument-based performance measure (e.g., PRO-PM), identify whether (and how) proxy responses are allowed.

N/A

S.16. Survey/Patient-reported data *(If measure is based on a survey or instrument, provide instructions for data collection and guidance on minimum response rate.)*

Specify calculation of response rates to be reported with performance measure results.

N/A

S.17. Data Source *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*

If other, please describe in S.18.

Claims, Electronic Health Data, Electronic Health Records, Other, Registry Data

S.18. Data Source or Collection Instrument *(Identify the specific data source/data collection instrument (e.g. name of database, clinical registry, collection instrument, etc., and describe how data are collected.)*

IF instrument-based, identify the specific instrument(s) and standard methods, modes, and languages of administration.

Although Get With The Guidelines (GWTG) – ACTION Registry (AR-G) data was used in the development of measure specification, this measure is not intended to be used as a registry measure. All data elements derived from the EHR have been tested for feasibility at multiple hospital sites, as shown in the Measure Testing Form section 2b1.3.

Data sources for the Medicare FFS measure:

1. Medicare Part A inpatient claims: This data source contains claims data for FFF inpatient and outpatient services and Medicare inpatient hospital care.
2. Medicare Enrollment Database (EDB): This database contains Medicare beneficiary demographic, benefit/coverage, and vital status information. This data source was used to obtain information on several inclusion/exclusion indicators such as Medicare status on admission.
3. Patients' electronic health records: The clinical data elements used in the risk models for this measure will be derived from patients EHRs. The measure was tested using data from EHRs.

S.19. Data Source or Collection Instrument *(available at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*

No data collection instrument provided

S.20. Level of Analysis *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*

Facility

S.21. Care Setting *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*

Inpatient/Hospital

If other:

S.22. COMPOSITE Performance Measure - Additional Specifications *(Use this section as needed for aggregation and weighting rules, or calculation of individual performance measures if not individually endorsed.)*

N/A

2. Validity – See attached Measure Testing Submission Form

HOY4_Hybrid_AMI_Mortality_Data_Dictionary_v1.0-636507676664095439.xls,NQF_Testing_Attachment_7.1_v2.0.docx

2.1 For maintenance of endorsement

Reliability testing: If testing of reliability of the measure score was not presented in prior submission(s), has reliability testing of the measure score been conducted? If yes, please provide results in the Testing attachment. Please use the most current version of the testing attachment (v7.1). Include information on all testing conducted (prior testing as well as any new testing); use red font to indicate updated testing.

Yes

2.2 For maintenance of endorsement

Has additional empirical validity testing of the measure score been conducted? If yes, please provide results in the Testing attachment. Please use the most current version of the testing attachment (v7.1). Include information on all testing conducted (prior testing as well as any new testing); use red font to indicate updated testing.

Yes

2.3 For maintenance of endorsement

Risk adjustment: For outcome, resource use, cost, and some process measures, risk-adjustment that includes social risk factors is not prohibited at present. Please update sections 1.8, 2a2, 2b1,2b4.3 and 2b5 in the Testing attachment and S.140 and S.11 in the online submission form. NOTE: These sections must be updated even if social risk factors are not included in the risk-adjustment strategy. You MUST use the most current version of the Testing Attachment (v7.1) -- older versions of the form will not have all required questions.

Yes - Updated information is included

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.
- 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] 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] 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]

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]

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; [14,15] 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] **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:	☒other: American Community Survey

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 <u>all</u> 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:	☐other:

1.5. How many and which <u>measured entities</u> 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 <u>patients</u> 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:

<u>For reliability testing</u>

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

☒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

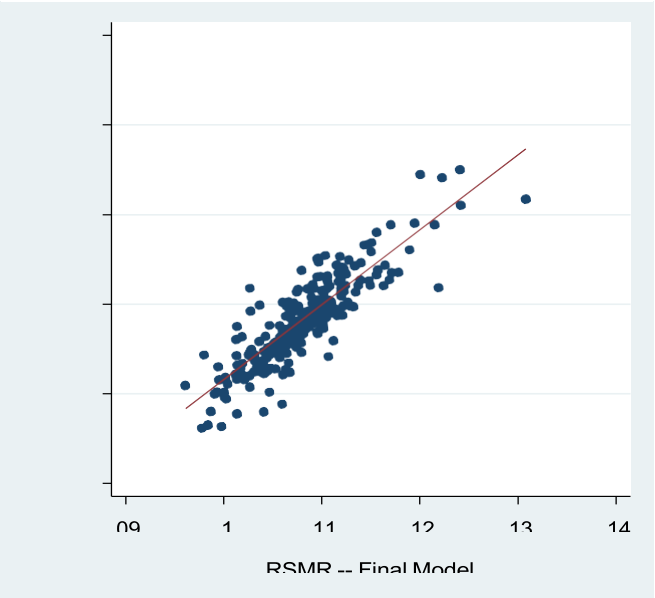
Table. Percent Agreement and Confidence Intervals: Comparison of EHR-Extracted and Manually Abstracted CCDE (Dataset 4, N=92)

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)



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

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 care institution	615	2.80%
3. Unknown death records with missing vital status) in Medicare Enrollment Database	0	0.0%
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.

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 risk categories**

☐**Other,**

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

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.

3. Feasibility

Extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

3a. Byproduct of Care Processes

For clinical measures, the required data elements are routinely generated and used during care delivery (e.g., blood pressure, lab test, diagnosis, medication order).

3a.1. Data Elements Generated as Byproduct of Care Processes.

Generated or collected by and used by healthcare personnel during the provision of care (e.g., blood pressure, lab value, diagnosis, depression score), Coded by someone other than person obtaining original information (e.g., DRG, ICD-9 codes on claims)

If other:

3b. Electronic Sources

The required data elements are available in electronic health records or other electronic sources. If the required data are not in electronic health records or existing electronic sources, a credible, near-term path to electronic collection is specified.

3b.1. To what extent are the specified data elements available electronically in defined fields (*i.e., data elements that are needed to compute the performance measure score are in defined, computer-readable fields*) Update this field for **maintenance of endorsement**.

ALL data elements are in defined fields in a combination of electronic sources

3b.2. If ALL the data elements needed to compute the performance measure score are not from electronic sources, specify a credible, near-term path to electronic capture, OR provide a rationale for using other than electronic sources. For **maintenance of endorsement**, if this measure is not an eMeasure (eCQM), please describe any efforts to develop an eMeasure (eCQM).

All data elements needed to compute performance measure score are captured electronically.

3b.3. If this is an eMeasure, provide a summary of the feasibility assessment in an attached file or make available at a measure-specific URL. Please also complete and attach the NQF Feasibility Score Card.

Attachment: nqf_ecqm_feasibility_scorecard_v1.0.xlsx

3c. Data Collection Strategy

Demonstration that the data collection strategy (e.g., source, timing, frequency, sampling, patient confidentiality, costs associated with fees/licensing of proprietary measures) can be implemented (e.g., already in operational use, or testing demonstrates that it is ready to put into operational use). For eMeasures, a feasibility assessment addresses the data elements and measure logic and demonstrates the eMeasure can be implemented or feasibility concerns can be adequately addressed.

3c.1. Required for maintenance of endorsement. Describe difficulties (as a result of testing and/or operational use of the measure) regarding data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, time and cost of data collection, other feasibility/implementation issues.

IF instrument-based, consider implications for both individuals providing data (patients, service recipients, respondents) and those whose performance is being measured.

During testing the measure specifications in five hospitals with various EHR systems, few difficulties were found. As shown in the NQF Testing Form, capture rate was high, and data element validity, or agreement between EHR data and chart data was high. We found that some initial time was required for a hospital to map the data elements in the measure specifications to their own EHR system. However, once these data elements are mapped, a hospital could submit many of these data elements for other hybrid measures, once implemented.

3c.2. Describe any fees, licensing, or other requirements to use any aspect of the measure as specified *(e.g., value/code set, risk model, programming code, algorithm)*.

There are no fees for use of this measure.

4. Usability and Use

Extent to which potential audiences (e.g., consumers, purchasers, providers, policy makers) are using or could use performance results for both accountability and performance improvement to achieve the goal of high-quality, efficient healthcare for individuals or populations.

4a. Accountability and Transparency

Performance results are used in at least one accountability application within three years after initial endorsement and are publicly reported within six years after initial endorsement (or the data on performance results are available). If not in use at the time of initial endorsement, then a credible plan for implementation within the specified timeframes is provided.

4.1. Current and Planned Use

NQF-endorsed measures are expected to be used in at least one accountability application within 3 years and publicly reported within 6 years of initial endorsement in addition to performance improvement.

Specific Plan for Use	Current Use (for current use provide URL)
Not in use	

4a1.1 For each CURRENT use, checked above (update for <u>maintenance of endorsement</u>), provide:

- Name of program and sponsor
- Purpose
- Geographic area and number and percentage of accountable entities and patients included
- Level of measurement and setting

N/A

4a1.2. If not currently publicly reported OR used in at least one other accountability application (e.g., payment program, certification, licensing) what are the reasons? (*e.g., Do policies or actions of the developer/steward or accountable entities restrict access to performance results or impede implementation?*)

This measure was proposed and finalized into the Center for Medicare and Medicaid Services Innovation (CMMI) Advancing Care Coordination Through Episode Payment Models (EPM) five-year bundled payment model in January 2017 (82 FR 180). However, in December 2017, CMMI finalized the cancellation of the bundled payment model that included the hybrid AMI mortality measure (82 FR 57066).

This measure was also signaled in the FY 2016 Hospital Inpatient Quality Reporting (HIQR) final rule in August 2015 (80 FR 49698).

4a1.3. If not currently publicly reported OR used in at least one other accountability application, provide a credible plan for implementation within the expected timeframes -- any accountability application within 3 years and publicly reported within 6 years of initial endorsement. (*Credible plan includes the specific program, purpose, intended audience, and timeline for implementing the measure within the specified timeframes. A plan for accountability applications addresses mechanisms for data aggregation and reporting.*)

This measure is final and ready to be implemented. This measure builds upon the work of the hybrid hospital-wide readmission (HWR) measure (NQF# 0230) currently implemented in the HIQR program as a voluntary measure for reporting. Although, the implementation plan for the hybrid AMI mortality measure has not yet been determined, hospitals will submit data in 2018 for the hybrid HWR measure. This hybrid AMI mortality measure, which uses nearly identical EHR-derived data elements, can be implemented by CMS in future regulation sand is suitable for the HIQR program, the Hospital Value Based Payment (HVBP) program, or a future EPM under a CMMI program.

4a2.1.1. Describe how performance results, data, and assistance with interpretation have been provided to those being measured or other users during development or implementation.

How many and which types of measured entities and/or others were included? If only a sample of measured entities were included, describe the full population and how the sample was selected.

N/A

4a2.1.2. Describe the process(es) involved, including when/how often results were provided, what data were provided, what educational/explanatory efforts were made, etc.

N/A

4a2.2.1. Summarize the feedback on measure performance and implementation from the measured entities and others described in 4d.1.

Describe how feedback was obtained.

N/A

4a2.2.2. Summarize the feedback obtained from those being measured.

N/A

4a2.2.3. Summarize the feedback obtained from other users

N/A

4a2.3. Describe how the feedback described in 4a2.2.1 has been considered when developing or revising the measure specifications or implementation, including whether the measure was modified and why or why not.

N/A

Improvement

Progress toward achieving the goal of high-quality, efficient healthcare for individuals or populations is demonstrated. If not in use for performance improvement at the time of initial endorsement, then a credible rationale describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.

4b1. Refer to data provided in 1b but do not repeat here. Discuss any progress on improvement (trends in performance results, number and percentage of people receiving high-quality healthcare; Geographic area and number and percentage of accountable entities and patients included.)

If no improvement was demonstrated, what are the reasons? If not in use for performance improvement at the time of initial endorsement, provide a credible rationale that describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.

Although this measure is not in public reporting, the harmonized claims-only AMI mortality measure has been in the HIQR program for many years, and has shown a slight decrease over time in AMI mortality. T Because it includes clinical information gathered and used in the course of patient care, the hybrid AMI mortality measure has improved credibility and face validity among stakeholders. It also aligns with CMS's goal to incorporate electronic clinical data into quality measures wherever possible.

4b2. Unintended Consequences

The benefits of the performance measure in facilitating progress toward achieving high-quality, efficient healthcare for individuals or populations outweigh evidence of unintended negative consequences to individuals or populations (if such evidence exists).

4b2.1. Please explain any unexpected findings (positive or negative) during implementation of this measure including unintended impacts on patients.

We did not identify any unintended consequences during measure development, model testing, or testing the risk variables in hospital settings.

4b2.2. Please explain any unexpected benefits from implementation of this measure.

N/A; measure not currently implemented.

5. Comparison to Related or Competing Measures

If a measure meets the above criteria and there are endorsed or new related measures (either the same measure focus or the same target population) or competing measures (both the same measure focus and the same target population), the measures are compared to address harmonization and/or selection of the best measure.

5. Relation to Other NQF-endorsed Measures

Are there related measures (conceptually, either same measure focus or target population) or competing measures (conceptually both the same measure focus and same target population)? If yes, list the NQF # and title of all related and/or competing measures.

Yes

5.1a. List of related or competing measures (selected from NQF-endorsed measures)

5.1b. If related or competing measures are not NQF endorsed please indicate measure title and steward.

N/A

5a. Harmonization of Related Measures

The measure specifications are harmonized with related measures;

OR

The differences in specifications are justified

5a.1. If this measure conceptually addresses EITHER the same measure focus OR the same target population as NQF-endorsed measure(s):

Are the measure specifications harmonized to the extent possible?

Yes

5a.2. If the measure specifications are not completely harmonized, identify the differences, rationale, and impact on interpretability and data collection burden.

The measure specifications are, by design, not completely harmonized in that the current measure uses clinical data elements collected from EHR for risk adjustment, and the measures listed above use claims data for risk adjustment. Additionally, the outcome in measure #0730 is inpatient mortality rather than 30-day mortality. Inpatient mortality rates can be influenced by hospital length of stay, so 30-day measures that establish a standard follow-up period are more appropriate for profiling a diverse group of hospitals (Drye et al., 2012). The measures listed above have target populations aged 18+, whereas the current measure's target population is age 65+. The exclusion criteria of the current measure are largely similar to those of measure #0230. Reference: Drye EE, Normand SL, Wang Y, Ross JS, Schreiner GC, Han L, Rapp M, Krumholz HM. Comparison of hospital risk-standardized mortality rates calculated by using in-hospital and 30-day models: an observational study with implications for hospital profiling. Ann Intern Med. 2012 Jan 3;156(1 Pt 1):19-26.

5b. Competing Measures

The measure is superior to competing measures (e.g., is a more valid or efficient way to measure);

OR

Multiple measures are justified.

5b.1. If this measure conceptually addresses both the same measure focus and the same target population as NQF-endorsed measure(s):

Describe why this measure is superior to competing measures (e.g., a more valid or efficient way to measure quality); OR provide a rationale for the additive value of endorsing an additional measure. (Provide analyses when possible.)

The use of clinical data elements that are measured in patients during the course of diagnosis and treatment have greater face validity among providers and produced a model with better discrimination (higher c-statistic) compared with the model risk- adjusted with claims data only. However, the hybrid measure has not yet been implemented and we recommend continues endorsement of both measures until such a time as CMS implements the hybrid measure.

Appendix

A.1 Supplemental materials may be provided in an appendix. All supplemental materials (such as data collection instrument or methodology reports) should be organized in one file with a table of contents or bookmarks. If material pertains to a specific submission form number, that should be indicated. Requested information should be provided in the submission form and required attachments. There is no guarantee that supplemental materials will be reviewed.

Available at measure-specific web page URL identified in S.1 Attachment:

Contact Information

Co.1 Measure Steward (Intellectual Property Owner): Centers for Medicare & Medicaid Services (CMS)

Co.2 Point of Contact: Lein, Han, Lein.han@cms.hhs.gov, 410-786-0205-

Co.3 Measure Developer if different from Measure Steward: Yale New Haven Health Services Corporation/Center for Outcomes Research and Evaluation (YNHHSC/CORE)

Co.4 Point of Contact: Karen, Dorsey, karen.dorsey@yale.edu, 203-764-5700-

Additional Information

Ad.1 Workgroup/Expert Panel involved in measure development

Provide a list of sponsoring organizations and workgroup/panel members' names and organizations. Describe the members' role in measure development.

The following experts provided insight and guidance during measure development.

American College of Cardiology, the National Cardiovascular Data Registry, and the Duke Clinical Research Institute:

Frederick Masoudi, MD, MSPH

Gregg Fonarow, MD

Joanne Foody, MD

James Jollis, MD

James DeLemos, MD, PhD

Joseph Drozda, MD

George Dengas, MD, PhD

Vernon Anderson, MD

Lara Slattery, MHS

Electronic Health Records (EHR) System Experts:

Adam Landman, MD

Christopher Mast, MD, MS

Venkatesh Janakiraman

Office of the National Coordinator for Health Information Technology (ONC):

Jacob Reider, MD

Lauren Richie, MA

Members of Sentara Healthcare, Kaiser Permanente, Veterans Health Affairs, Mid America Heart Institute, Duke Clinical Research Institute, and Statewide Planning and Resource Cooperative System (SPARCS):

Gabriel Escobar, MD

John Brush, MD

John Parker, MD

Marta Render, MD

David Magid, MD

Edward Hannan, PhD, MS, MS

John Spertus, MD, MPH

Mikhail Kosiborod, MD

James Tcheng, MD

We would like to acknowledge Jeremy Michel, MD, Postdoctoral Fellow at the Yale Center for Medical Informatics, for his valuable input during the eSpecification process.

Additionally, researchers at Abt Associates and their subcontractors eSpecified and tested the eMeasure in collaboration with the CORE team.

Measure Developer/Steward Updates and Ongoing Maintenance

Ad.2 Year the measure was first released: 2016

Ad.3 Month and Year of most recent revision: 09, 2016

Ad.4 What is your frequency for review/update of this measure? Annual

Ad.5 When is the next scheduled review/update for this measure? 04, 2018

Ad.6 Copyright statement: N/A

Ad.7 Disclaimers: N/A

Ad.8 Additional Information/Comments: N/A