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 #: 2558

Measure Title: Hospital 30-Day, All-Cause, Risk-Standardized Mortality Rate (RSMR) Following Coronary Artery Bypass Graft (CABG) Surgery

Measure Steward: Centers for Medicare & Medicaid Services

Brief Description of Measure: The measure estimates a hospital-level, risk-standardized mortality rate (RSMR) for patients discharged from the hospital following a qualifying isolated CABG procedure. Mortality is defined as death from any cause within 30 days of the procedure date of an index CABG admission. An index CABG admission is the hospitalization for a qualifying isolated CABG procedure considered for the mortality outcome. The measure was developed using Medicare Fee-for-Service (FFS) patients 65 years and older and was tested in all-payer patients 18 years and older.

Developer Rationale: The goal of this measure is to improve patient outcomes by providing patients, physicians, hospitals, and policy makers with information about hospital-level, risk-standardized mortality rates following hospitalization for a qualifying isolated CABG procedure. 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 would be expected based on each institution's patient case mix, and therefore promote hospital quality improvement and better inform consumers about care quality.

CABG is a priority area for outcomes measure development because it is a common procedure associated with considerable morbidity, mortality, and health care spending. Between 2013 and 2016, there were 138,785 hospitalizations for CABG surgery among Medicare FFS patients in the U.S [1].

CABG surgeries are costly procedures that account for the majority of major cardiac surgeries performed nationally. In fiscal year 2014, isolated CABG surgeries accounted for almost half (40.59%) of all cardiac surgery hospital admissions in Massachusetts [2]. In 2014, the average Medicare payment was $32,499 for CABG without valve and $45,873 for CABG plus valve surgeries [3].

1. Simoes J, Grady J, DeBuhr J, et al. 2017 Procedure-Specific Measures Updates and Specifications Report Hospital-Level 30-Day Risk-Standardized Mortality Measures. http://www.qualitynet.org/dcs/ContentServer?c=Page&pagename=QnetPublic/Page/QnetTier3&cid=1163010421830. Accessed March 23, 2018.

2. Massachusetts Data Analysis Center. Adult Coronary Artery Bypass Graft Surgery in the Commonwealth of Massachusetts. https://www.mass.gov/files/documents/2017/12/14/cabg-fy2014.pdf. Accessed March 23, 2018.

3. Pennsylvania Health Care Cost Containment Council. Hospital Medicare Payment. http://www.phc4.org/reports/cabg/16/docs/Hospital%20Medicare%20Payment.pdf. Accessed March 23, 2018

Numerator Statement: The outcome for this measure is 30-day all-cause mortality. Mortality is defined as death for any reason within 30 days of the procedure date from the index admission for patients 18 and older discharged from the hospital after undergoing isolated CABG surgery.

Denominator Statement: This claims-based measure can be used in either of two patient cohorts: (1) patients aged 65 years or older or (2) patients aged 18 years or older. We have tested the measure in both age groups.

The cohort includes admissions for patients who receive a qualifying isolated CABG procedure (see the attached Data Dictionary) and with a complete claims history for the 12 months prior to admission. CMS publicly reports this measure for those patients 65 years or older who are Medicare FFS beneficiaries admitted to non-federal hospitals.

Denominator Exclusions: The CABG surgery mortality measure excludes index admissions for patients:

1. With inconsistent or unknown vital status or other unreliable demographic (age and gender) data; or,
2. Discharged against medical advice (AMA).

For patients with more than one qualifying CABG surgery admission in the measurement period, the first CABG admission is selected for inclusion in the measure and the subsequent CABG admission(s) are excluded from the cohort.

Measure Type: Outcome

Data Source: Claims

Level of Analysis: Facility

IF Endorsement Maintenance – Original Endorsement Date: Most Recent Endorsement Date:

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:**

- This outcome measure is based on evidence that aspects of perioperative, intra and peri-operative, and post-operative care practices can reduce 30-day mortality rates following coronary artery bypass graft (CABG) surgery.

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 provided performance data from 1,185 hospitals and 138,661 admissions from July 1, 2013 to June 30, 2016. Reported hospital-level risk-standardized mortality rate was 3.3%, ranging from 1.3% - 7.4%.
- *Empirical data* demonstrating a relationship between the outcome to at least one healthcare process is now required. NQF guidance states that a 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.

Question for the Committee:

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

Guidance from the Evidence Algorithm: Measure assesses a health outcome (Box 1) → The relationship between the outcome and the intervention demonstrated by performance 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 performance data from 1,185 hospitals and 138,661 admissions from July 1, 2013 to June 30, 2016. Reported mean hospital-level risk-standardized mortality rate was 3.3%, ranging from 1.3% - 7.4%. The full table is available here.

	July 2013 – June 2014	July 2014 – June 2015	July 2015 – June 2016	July 2013-June 2016
# hospitals	1,158	1,150	1,151	1,185
# admissions	46,279	46,123	46,259	138,1661
Mean (SD)	3.2 (0.5)	3.4 (0.7)	3.2 (0.7)	3.3 (0.9)
Range	1.9-6.0	1.8-6.7	1.4-6.8	1.3-7.4
50th percentile	3.1	3.2	3.0	3.1

Disparities

- The developer provided performance data for July 2013 – June 2016 by proportion of dual eligible patients, African-American patients, and by the proportion of patients with AHRQ SES Index Scores equal to or below 42.6. Median scores were higher in hospitals with higher proportions of dual eligible patients and of patients with SES index scores.

	Low proportion of Dual Eligible	**High proportion of Dual Eligible**	**Low proportion of AA**	**High proportion of AA**	**Low proportion ≤42.6**	**High proportion ≤42.6**
# entities	260	260	259	266	259	259
# patients	123,442	13,628	131,354	7,307	112,666	25,995
Maximum	7.4	7.2	6.0	6.4	7.4	7.2
90th	4.2	4.8	4.5	4.6	4.1	4.8
75th	3.5	3.9	3.8	3.8	3.4	4.1
Median	3.1	3.2	3.2	3.1	2.9	3.5
25th	2.6	2.7	2.7	2.7	2.5	2.8
10th	2.3	2.4	2.4	2.3	2.1	2.4
Minimum	1.3	1.5	1.6	1.5	1.3	1.7

Questions for the Committee:

- *Is there a gap in care in mortality following CABG that warrants a national performance measure?*

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)

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.

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

Evaluators: Jennifer Perloff, Larry Glance, Jeff Geppert

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

Questions for the Committee regarding reliability:

- o *Do you have any concerns that the measure can be consistently implemented (i.e., are measure specifications adequate)?*
- o *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:

- o *Do you have any concerns regarding the validity of the measure (e.g., exclusions, risk-adjustment approach, etc.)?*
- o *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**

Evaluation A: Scientific Acceptability

Measure Number: 2558

Measure Title: Hospital 30-Day, All-Cause, Risk-Standardized Mortality Rate (RSMR) Following Coronary Artery Bypass Graft (CABG) Surgery

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.

   Comment: I wasn't sure if these two statements were conflicting:

   Denominator: If a patient has more than one qualifying isolated CABG admission in a year, one hospitalization is randomly selected for inclusion in the measure

   Exclusions: For patients with more than one qualifying CABG surgery admission in the measurement period, the first CABG admission is selected for inclusion in the measure and the subsequent CABG admission(s) are excluded from the cohort

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)

Comment: There is no data reported on the distribution of the reliability statistic; only the average or median is reported.

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]

    Comment: There is no data reported on the distribution of the reliability statistic; only the average or median is reported.

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

    Comment: SES factors were evaluated and ultimately determined not to be material to model performance

    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)

Comment: The testing documentation indicates that there is both an administrative data and registry specification, although no testing data on the comparability of these two data sources or methods is provided.

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 <u>empirical</u> 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)

Comment: There is no empirical validity testing of the performance score, in the sense of a stated hypothesis and demonstration that the performance score is aligned with the quality construct. The validity of the risk adjustment model is a sensitivity, specificity type analysis, which is usually relevant to the person level validity.

18. Was validity testing conducted with <u>computed performance measure scores</u> 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.]

Comment: For a measure that has been around this long, there should be an empirically testing hypothesis. Ideally the hypothesis would be that the measure is aligned with a quality construct (systematic and persistent behavior casually related to better patient outcomes) and the empirical testing would demonstrate the relationship between the measure and that construct (e.g. either a direct measure of the construct, or an outcome or process related to the same construct).

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)

Evaluation B for Scientific Acceptability

Measure Number: 2258

Measure Title: Hospital 30-Day, All-Cause, Risk-Standardized Mortality Rate (RSMR) Following Coronary Artery Bypass Graft (CABG) Surgery

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 <u>VALIDITY</u> testing** of <u>patient-level data</u> 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 <u>ONLY</u> 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 <u>data element validity testing</u> – 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 <u>all</u> testing results:

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

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

    ☐Low (please explain below) [NOTE: Should rate <u>LOW</u> 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 <u>not</u> 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)

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 this as moderate because the administrative model tended to underestimate RSMRs in poor-performing hospitals compared to the clinical model, and higher RSMRs in high-performing hospitals.

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)

Evaluation C for Scientific Acceptability

Measure Number: 2558

Measure Title: Hospital 30-Day, All-Cause, Risk-Standardized Mortality Rate (RSMR) Following Coronary Artery Bypass Graft (CABG) Surgery

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)

The authors report an ICC of 0.35 for the test/re-test reliability, which seems relatively low for a CABG specific readmissions measure where the event of interest is discrete and easily observable. The authors point out that the ICC is a conservative measure of reliability and that risk adjusted measures are 'complex constructs'. Pointing to the importance of context, they go on to show that latent constructs tend to have lower reliability. The authors present lots on information on the reliability of similar measures – this seems somewhat tangential since the question is not what have others accepted, but what can providers tolerate as reasonable. In using a second method, the authors report a 'unit' reliability of 0.851. All told, moderate seems reasonable, although somewhat arbitrary.

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 <u>VALIDITY</u> testing** of <u>patient-level data</u> 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 <u>ONLY</u> 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 <u>data element validity testing</u> – 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 <u>all</u> testing results:

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

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

    ☐Low (please explain below) [NOTE: Should rate <u>LOW</u> 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 <u>not</u> required, but check with NQF staff]

See comments under 5 above.

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)

The model is data driven, using 1000 replications and a cut off of 70% for inclusion in the model (the risk factor shows up in 70% or more of the replications of the model). As a second stage the nominated predictors undergo clinical review. Some factors are forced into the model based on theoretical importance. This is an excellent way to identify factors from a large pool of potential risk factors.

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)

It looks like all of the validation work took place at the hospital level. I'm not sure why the measure Form says 'hospital/facility/agency'. I particularly like the use of the New York Registry data to validate the risk models. Much of the validation focuses on the function of the risk models – this makes sense to some extent because it is a more 'abstract' part of the measure. The validation data was 2008-2011 which pre-dates the conversion to ICD-10. The conversion of the measure sounds reasonable. However, changes in coding practices could have an impact on reliability and validity.

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 <u>patient-level data elements</u>?

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 <u>not</u> been previously endorsed; otherwise, skip Questions #24-25 and go to Question #26)

24. Was <u>face validity</u> 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)

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 reports that data elements are coded by someone other than the person obtaining the original information and that all data elements are in defined fields in electronic claims.
- The developer also reports that administrative data are routinely collected during the billing process and no additional data collection is required.
- The developer reports there are no fees associated with the measure.

Questions for the Committee:

o *Are the required data elements routinely generated and used during care delivery?*
o *Are the required data elements available in electronic form, e.g., EHR or other electronic sources?*

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

Committee Pre-evaluation Comments: Criteria 3: Feasibility

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**

Accountability program details

- This measure is used in Hospital Inpatient Quality Reporting (IQR), and has been finalized for the Hospital Value-Based Purchasing (VBP) program. Under the Hospital IQR, CMS collects data from hospitals paid through the Inpatient Prospective Payment System and publicly displays this data to help consumers make informed decisions about health care. The Hospital VBP is a payment program that encourages hospitals to improve the quality and safety of care for Medicare beneficiaries and all patients during inpatient stays.

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 reports that in April of each year, hospitals have access to various resources related to the measure. In July of each year, measure results are posted on Hospital Compare (a tool to compare quality of care among hospitals).
- Hospitals can submit questions or comments about the measure and these questions/comments are considered during the measure reevaluation process.

- Additionally, the developer reports they routinely complete literature reviews for research related to this measure.

Questions for the Committee:

- o *How have the performance results be used to further the goal of high-quality, efficient healthcare?*
- o *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 developer reports that the median risk-standardized mortality rate decreased by 0.1 absolute percentage points from July 2013-June 2014 (median – 3.1%) to July 2015-June 2016 (median – 3.0%).

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

- The developer indicated this question was not applicable.

Potential harms

- The developer indicated this question was not applicable.

Questions for the Committee:

- o *How can the performance results be used to further the goal of high-quality, efficient healthcare?*

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

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

Criterion 5: Related and Competing Measures

Related or competing measures

0114 : Risk-Adjusted Postoperative Renal Failure

0115 : Risk-Adjusted Surgical Re-exploration

0119 : Risk-Adjusted Operative Mortality for CABG

0122 : Risk-Adjusted Operative Mortality for Mitral Valve (MV) Replacement + CABG Surgery

0123 : Risk-Adjusted Operative Mortality for Aortic Valve Replacement (AVR) + CABG Surgery

0129 : Risk-Adjusted Postoperative Prolonged Intubation (Ventilation)

0130 : Risk-Adjusted Deep Sternal Wound Infection

0131 : Risk-Adjusted Stroke/Cerebrovascular Accident

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

0468 : Hospital 30-day, all-cause, risk-standardized mortality rate (RSMR) following pneumonia hospitalization

0535 : 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

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

1502 : Risk-Adjusted Operative Mortality for Mitral Valve (MV) Repair + CABG Surgery

1893 : Hospital 30-Day, all-cause, risk-standardized mortality rate (RSMR) following chronic obstructive pulmonary disease (COPD) hospitalization

2515 : Hospital 30-day, all-cause, unplanned, risk-standardized readmission rate (RSRR) following coronary artery bypass graft (CABG) surgery

Harmonization

- The developer reports that NQF #2558 has the same target population and measure focus as NQF #0119 Risk Adjusted Operative Mortality for CABG and that they have sought to harmonize with #0119. “The potential sources of discrepancy are target patient population, age, isolated CABG, period of observation, and included hospitals. The STS measure also assesses both deaths occurring during CABG hospitalization (in-hospital death, even if after 30 days) and deaths occurring within 30 days of procedure date”.

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

Public and Member Comments

Comments and Member Support/Non-Support Submitted as of: June 19, 2018

- Medtronic appreciates the opportunity to submit comments to the National Quality Forum's Surgery Portfolio Committee on the Spring 2018 Cycle Measures. Medtronic supports efforts to "alleviate pain, restore health, and extend life" and Medtronic's Minimally Invasive Therapies Group is actively engaged in developing innovative solutions for monitoring and patient safety to assist in the early detection of preventable, adverse events. We commend the committee for their thorough review and support continued endorsement of these measures.
- The Federation of American Hospitals (FAH) appreciates the opportunity to comment on Measure #2558: Hospital 30-Day, All-Cause, Risk-Standardized Mortality Rate (RSMR) Following Coronary Artery Bypass Graft (CABG) Surgery. The FAH identified several questions and concerns that we note for the Standing Committee’s consideration including:

1A. Evidence to Support the Measure Focus: The FAH does not disagree with the importance of assessing the mortality rates of those patients who had a hospital admission. However, the FAH does not believe that the Center for Medicare and Medicaid Services (CMS) has provided sufficient evidence for this measure and other mortality measures included in CMS programs that a death in the 30 days following an inpatient admission is a predictor of the quality of care provided by a hospital and may well be due to other factors outside of a hospital’s control. The FAH does not believe that adequate justification has been provided for selection of a 30-day window. On review of the evidence provided for this measure, most, if not all, of the studies cited focus on surgical technique and intra-operative interventions and we did not identify any evidence to support measuring mortality using a 30-day time period.

2B. Validity: The FAH questions whether the measure meets the requirements for validity testing for measures undergoing maintenance given the lack of empirical validity testing. Only testing for face validity and the validity of the risk adjustment model were provided.

The FAH would like to again reiterate our disappointment in the minimal set of variables used to test whether social risk factors should be included in the risk adjustment model. As experience is gained and additional factors are available related to the community in which the patient resides such as access to transportation or pharmacies, we hope to see further analysis and testing be completed in the near future.

The FAH would also note that testing of social risk factors in the risk adjustment model demonstrated a statistically significant association for each of the two variables; yet, the developer determined that their inclusion was not needed given the lack of improvement of model performance and hospital profiling. Given the minimal variation in performance scores for this measure, which in 2016 ranged from 1.3% to 7.4%, FAH is concerned that what may appear as small changes in performance scores when either of the two variables are included could shift a hospital's risk-standardized mortality rate (RSMR) (e.g., from worse than the national rate to no different than the national rate). Regrettably, this analysis was not provided and would provide useful information in determining whether inclusion of these risk factors is warranted.

In addition, the FAH is concerned that there is insufficient variation in performance across hospitals to support this measure's use in accountability programs. Specifically, the performance scores reported in 2b4. Identification of Statistically Significant and Meaningful Difference in Performance are generally low with only 17 hospitals identified as better than the national rate, 1,004 as no different than the national rate, and 18 as worse than the national rate.

- **Of the 1 NQF member who have submitted a support/non-support choice:**
  - 1 supports the measure

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 #: 2558

Corresponding Measures:

De.2. **Measure Title:** Hospital 30-Day, All-Cause, Risk-Standardized Mortality Rate (RSMR) Following Coronary Artery Bypass Graft (CABG) Surgery

Co.1.1. **Measure Steward:** Centers for Medicare & Medicaid Services

De.3. Brief Description of Measure: The measure estimates a hospital-level, risk-standardized mortality rate (RSMR) for patients discharged from the hospital following a qualifying isolated CABG procedure. Mortality is defined as death from any cause within 30 days of the procedure date of an index CABG admission. An index CABG admission is the hospitalization for a qualifying isolated CABG procedure considered for the mortality outcome. The measure was developed using Medicare Fee-for-Service (FFS) patients 65 years and older and was tested in all-payer patients 18 years and older.

1b.1. Developer Rationale: The goal of this measure is to improve patient outcomes by providing patients, physicians, hospitals, and policy makers with information about hospital-level, risk-standardized mortality rates following hospitalization for a qualifying isolated CABG procedure. 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 would be expected based on each institution's patient case mix, and therefore promote hospital quality improvement and better inform consumers about care quality.

CABG is a priority area for outcomes measure development because it is a common procedure associated with considerable morbidity, mortality, and health care spending. Between 2013 and 2016, there were 138,785 hospitalizations for CABG surgery among Medicare FFS patients in the U.S [1].

CABG surgeries are costly procedures that account for the majority of major cardiac surgeries performed nationally. In fiscal year 2014, isolated CABG surgeries accounted for almost half (40.59%) of all cardiac surgery hospital admissions in Massachusetts [2]. In 2014, the average Medicare payment was $32,499 for CABG without valve and $45,873 for CABG plus valve surgeries [3].

1. Simoes J, Grady J, DeBuhr J, et al. 2017 Procedure-Specific Measures Updates and Specifications Report Hospital-Level 30-Day Risk-Standardized Mortality Measures. http://www.qualitynet.org/dcs/ContentServer?c=Page&pagename=QnetPublic/Page/QnetTier3&cid=1163010421830. Accessed March 23, 2018.

2. Massachusetts Data Analysis Center. Adult Coronary Artery Bypass Graft Surgery in the Commonwealth of Massachusetts. https://www.mass.gov/files/documents/2017/12/14/cabg-fy2014.pdf. Accessed March 23, 2018.

3. Pennsylvania Health Care Cost Containment Council. Hospital Medicare Payment. http://www.phc4.org/reports/cabg/16/docs/Hospital%20Medicare%20Payment.pdf. Accessed March 23, 2018.

S.4. Numerator Statement: The outcome for this measure is 30-day all-cause mortality. Mortality is defined as death for any reason within 30 days of the procedure date from the index admission for patients 18 and older discharged from the hospital after undergoing isolated CABG surgery.

S.6. Denominator Statement: This claims-based measure can be used in either of two patient cohorts: (1) patients aged 65 years or older or (2) patients aged 18 years or older. We have tested the measure in both age groups.

The cohort includes admissions for patients who receive a qualifying isolated CABG procedure (see the attached Data Dictionary) and with a complete claims history for the 12 months prior to admission. CMS publicly reports this measure for those patients 65 years or older who are Medicare FFS beneficiaries admitted to non-federal hospitals.

If a patient has more than one qualifying isolated CABG admission in a year, the first CABG admission is selected for inclusion in the measure and the subsequent CABG admission(s) are excluded from the cohort.

S.8. Denominator Exclusions: The CABG surgery mortality measure excludes index admissions for patients:

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

2. Discharged against medical advice (AMA).

For patients with more than one qualifying CABG surgery admission in the measurement period, the first CABG admission is selected for inclusion in the measure and the subsequent CABG admission(s) are excluded from the cohort.

De.1. Measure Type: Outcome

S.17. Data Source: Claims

S.20. Level of Analysis: Facility

IF Endorsement Maintenance – Original Endorsement Date: Nov 12, 2014 **Most Recent Endorsement Date:** Nov 12, 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

Del18bHOY4CABGMortalityEndorsementMaintenanceEvidenceAttachment04022018.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.

Yes

1a Evidence (subcriterion 1a)

Measure Number (*if previously endorsed*)**:** 2558

Measure Title: Hospital 30-Day, All-Cause, Risk-Standardized Mortality Rate (RSMR) Following Coronary Artery Bypass Graft (CABG) Surgery

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

Date of Submission: Click here to enter a date

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*):

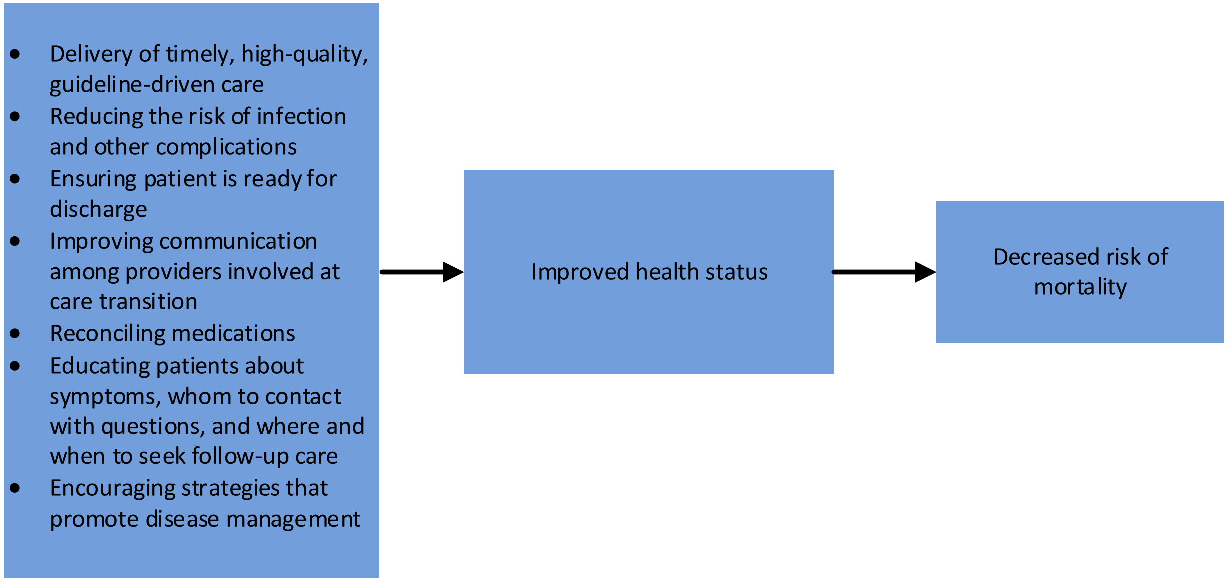
☐ Process:

☐ 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.



- Delivery of timely, high-quality, guideline-driven care
- Reducing the risk of infection and other complications
- Ensuring patient is ready for discharge
- Improving communication among providers involved at care transition
- Reconciling medications
- Educating patients about symptoms, whom to contact with questions, and where and when to seek follow-up care
- Encouraging strategies that promote disease management
Improved health status
Decreased risk of mortality

This is an outcome measure; the measure focus is all-cause mortality following isolated CABG procedures.

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 isolated CABG procedures. 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. 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 what would be expected based on their patient case-mix, and therefore promote hospital quality improvement and better inform consumers about quality of 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.)

N/A. This measure is not derived from patient report.

****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.

Mortality is the primary negative outcome associated with a surgical procedure. Many aspects of peri-operative care, intra- and peri-operative practices and several aspects of post-operative care, including prevention of and response to complications and coordinated transitions to the outpatient environment, have been shown to impact CABG mortality. A number of recent studies have demonstrated that improvements in care can reduce 30-day mortality rates.

References:

Emerson DA, Hynes CF, Greenberg MD, Trachiotis GD. Coronary Artery Bypass Grafting During Acute Coronary Syndrome: Outcomes and Comparison of Off-Pump to Conventional Coronary Artery Bypass Grafting at a Veteran Affairs Hospital. Innovations (Phila). 2015;10(3):157-162.

Johnson SH, Theurer PF, Bell GF, Laresca L, Leyden T, Prager RL. A statewide quality collaborative for process improvement: internal mammary artery utilization. Ann Thorac Surg. 2010; 90: 1158– 1164.

Kurlansky PA, Traad EA, Dorman MJ, Galbut DL, Ebra G. Bilateral Versus Single Internal Mammary Artery Grafting in the Elderly: Long-Term Survival Benefit. The Annals of thoracic surgery. 2015;100(4):1374-1381; discussion 1381-1372.

New York Department of Health (NYDH) https://www.health.ny.gov/statistics/diseases/cardiovascular/heart_disease/docs/2013-2015_adult_cardiac_surgery.pdf. March 7, 2018.

Northern New England Cardiovascular Disease Study Group. http://www.nnecdsg.org/pub_lit_2.htm. Accessed March 7, 2018.

Shroyer AL, Grover FL, Hattler B, et. al. On-pump versus off-pump coronary artery bypass surgery. N Engl J Med. 2009 Nov 5;361(19):1827-37.

Tranbaugh RF, Lucido DJ, Dimitrova KR, et al. Multiple arterial bypass grafting should be routine. *J Thorac Cardiovasc Surg.* 2015;150(6):1537-1544; discussion 1544-1535.

Williams JB, DeLong ER, Peterson ED, Dokholyan RS, Ou FS, Ferguson TB Jr.; Society of Thoracic Surgeons and the National Cardiac Database. Secondary prevention after coronary artery bypass graft surgery: findings of a national randomized controlled trial and sustained society-led incorporation into practice. Circulation. 2011; 123: 39– 45.

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 <u>systematic review of the body of evidence</u> 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

N/A. This measure is not an intermediate outcome, process, or structure performance measure.

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, hospitals, and policy makers with information about hospital-level, risk-standardized mortality rates following hospitalization for a qualifying isolated CABG procedure. 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 would be expected based on each institution's patient case mix, and therefore promote hospital quality improvement and better inform consumers about care quality.

CABG is a priority area for outcomes measure development because it is a common procedure associated with considerable morbidity, mortality, and health care spending. Between 2013 and 2016, there were 138,785 hospitalizations for CABG surgery among Medicare FFS patients in the U.S [1].

CABG surgeries are costly procedures that account for the majority of major cardiac surgeries performed nationally. In fiscal year 2014, isolated CABG surgeries accounted for almost half (40.59%) of all cardiac surgery hospital admissions in Massachusetts [2]. In 2014, the average Medicare payment was $32,499 for CABG without valve and $45,873 for CABG plus valve surgeries [3].

1. Simoes J, Grady J, DeBuhr J, et al. 2017 Procedure-Specific Measures Updates and Specifications Report Hospital-Level 30-Day Risk-Standardized Mortality Measures. http://www.qualitynet.org/dcs/ContentServer?c=Page&pagename=QnetPublic/Page/QnetTier3&cid=1163010421830. Accessed March 23, 2018.

2. Massachusetts Data Analysis Center. Adult Coronary Artery Bypass Graft Surgery in the Commonwealth of Massachusetts. https://www.mass.gov/files/documents/2017/12/14/cabg-fy2014.pdf. Accessed March 23, 2018.

3. Pennsylvania Health Care Cost Containment Council. Hospital Medicare Payment. http://www.phc4.org/reports/cabg/16/docs/Hospital%20Medicare%20Payment.pdf. Accessed March 23, 2018.

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.*

Variation in mortality rates indicates opportunity for improvement. We conducted analyses using data from July 1, 2013 to June 30, 2016 Medicare claims data (n= 138,66 admissions from 1,185 hospitals) and reported hospital-level RSRMs having a mean of 3.3% (SD=0.9) and a range of 1.3% - 7.4%. The median RSRR is 3.1% (20th and 70th percentiles are 2.6% and 3.6%, respectively). The distribution of RSRMs across hospitals is shown below:

Distribution of Hospital CABG RSMRs over Different Time Periods

Results for each data year

Characteristic	07/2013-06/2014	07/2014-06/2015	07/2015-06/2016	07-2013-06/2016
Number of Hospitals	1,158	1,150	1,151	1,185
Number of Admissions	46,279	46,123	46,259	138,661
Mean (SD)	3.2 (0.5)	3.4 (0.7)	3.2 (0.7)	3.3 (0.9)
Range (min. – max.)	1.9 – 6.0	1.8 – 6.7	1.4 – 6.8	1.3 – 7.4
Minimum	1.9	1.8	1.4	1.3
10th percentile	2.7	2.6	2.5	2.3
20th percentile	2.8	2.9	2.7	2.6
30th percentile	2.9	3.0	2.8	2.8
40th percentile	3.0	3.1	2.9	3.0
50th percentile	3.1	3.2	3.0	3.1
60th percentile	3.2	3.3	3.1	3.3
70th percentile	3.3	3.6	3.4	3.6
80th percentile	3.5	3.8	3.7	3.9
90th percentile	3.8	4.2	4.1	4.4
Maximum	6.0	6.7	6.8	7.4

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.

N/A

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.*

Distribution of CABG RSMRs by Proportion of Dual Eligible Patients:

Data Source: Medicare FFS claims

Dates of Data: July 2013 through June 2016

Characteristic	Hospitals with a low proportion (=5.6%) Dual Eligible patients	Hospitals with a high proportion (=13.4%) Dual Eligible patients
Number of Measured Entities (Hospitals)	260	260
Number of Patients	123,442 patients in low-proportion hospitals	13,628 in high-proportion hospitals
Maximum	7.4	7.2

Characteristic	Hospitals with a low proportion (=5.6%) Dual Eligible patients	Hospitals with a high proportion (=13.4%) Dual Eligible patients
90th percentile	4.2	4.8
75th percentile	3.5	3.9
Median (50th percentile)	3.1	3.2
25th percentile	2.6	2.7
10th percentile	2.3	2.4
Minimum	1.3	1.5

Distribution of CABG RSMRs by Proportion of African-American Patients:

Data Source: Medicare FFS claims

Dates of Data: July 2013 through June 2016

Characteristic	Hospitals with a low Proportion (=0.7%) African-American patients	Hospitals with a high proportion (=7.1%) African-American patients
Number of Measured Entities (Hospitals)	259	266
Number of Patients	131,354 patients in low-proportion hospitals	7,307 in high-proportion hospitals
Maximum	6.0	6.4
90th percentile	4.5	4.6
75th percentile	3.8	3.8
Median (50%)	3.2	3.1
25th percentile	2.7	2.7
10th percentile	2.4	2.3
Minimum	1.6	1.5

Distribution of CABG RSMRs by Proportion of Patients with AHRQ SES Index Scores Equal to or Below 42.6:

Data Source: Medicare FFS claims and The American Community Survey (2008-2012) data

Dates of Data: July 2013 through June 2016

Characteristic	Hospitals with low proportion of patients with AHRQ SES index score equal to or below 42.6 (=8.8%)	Hospitals with high proportion of patients with AHRQ SES index score equal to or below 42.6 (=26.8%)
Number of Measured Entities (Hospitals)	259	259

Characteristic	Hospitals with low proportion of patients with AHRQ SES index score equal to or below 42.6 (=8.8%)	Hospitals with high proportion of patients with AHRQ SES index score equal to or below 42.6 (=26.8%)
Number of Patients	112,666 patients in hospitals with low proportion of patients with AHRQ SES index score equal to or below 42.6	25,995 patients in hospitals with high proportion of patients with AHRQ SES index score equal to or below 42.6
Maximum	7.4	7.2
90th percentile	4.1	4.8
75th percentile	3.4	4.1
Median (50th percentile)	2.9	3.5
25th percentile	2.5	2.8
10th percentile	2.1	2.4
Minimum	1.3	1.7

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, Surgery : Cardiac Surgery

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

Care Coordination, Safety : Complications, Safety : Healthcare Associated Infections

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

Elderly

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/Measure-Methodology.html

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:** NQF_2558_CABG_Mortality_Data_Dictionary_12-30-16_v1.0.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.

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 for this measure is 30-day all-cause mortality. Mortality is defined as death for any reason within 30 days of the procedure date from the index admission for patients 18 and older discharged from the hospital after undergoing isolated CABG surgery.

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).

In the current publicly reported measure, we identify deaths for Medicare FFS patients 65 years or older in the Medicare Enrollment Database (EDB).

Outcome Attribution:

Attribution of the outcome in situations where a patient has multiple contiguous admissions, at least one of which involves a qualifying isolated CABG procedure is as follows:

1) If a patient undergoes a CABG procedure in the first hospital and is then transferred to a second hospital where there is no CABG procedure, the mortality outcome is attributed to the first hospital performing the index CABG procedure and the 30-day window starts with the date of index CABG procedure.

Rationale: A transfer following CABG is most likely due to a complication of the index procedure and that care provided by the hospital performing the CABG procedure likely dominates mortality risk even among transferred patients.

2) If a patient is admitted to a first hospital but does not receive a CABG procedure there and is then transferred to a second hospital where a CABG is performed, the mortality outcome is attributed to the second hospital performing the index CABG procedure and the 30-day window starts with the date of index CABG procedure.

Rationale: Care provided by the hospital performing the CABG procedure likely dominates mortality risk.

3) If a patient undergoes a CABG procedure in the first hospital and is transferred to a second hospital where another CABG procedure is performed, the mortality outcome is attributed to the first hospital performing the index (first) CABG procedure and the 30-day window starts with the date of index CABG procedure.

Rationale: A transfer following CABG is most likely due to a complication of the index procedure, and care provided by the hospital performing the index CABG procedure likely dominates mortality risk even among transferred patients.

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

This claims-based measure can be used in either of two patient cohorts: (1) patients aged 65 years or older or (2) patients aged 18 years or older. We have tested the measure in both age groups.

The cohort includes admissions for patients who receive a qualifying isolated CABG procedure (see the attached Data Dictionary) and with a complete claims history for the 12 months prior to admission. CMS publicly reports this measure for those patients 65 years or older who are Medicare FFS beneficiaries admitted to non-federal hospitals.

If a patient has more than one qualifying isolated CABG admission in a year, the first CABG admission is selected for inclusion in the measure and the subsequent CABG admission(s) are excluded from the cohort.

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.)*

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

The measure included index admissions for patients:

1. Having a qualifying isolated CABG surgery during the index admission;
2. Enrolled in Medicare fee-for-service (FFS) Part A and Part B for the 12 months prior to the date of the index admission, and enrolled in Part A during the index admission; and,
3. Aged 65 or over.

Isolated CABG surgeries are defined as those CABG procedures performed without the following concomitant valve or other major cardiac, vascular, or thoracic procedures:

- Valve procedures;
- Atrial and/or ventricular septal defects;
- Congenital anomalies;
- Other open cardiac procedures;
- Heart transplants;
- Aorta or other non-cardiac arterial bypass procedures;
- Head, neck, intracranial vascular procedures; or,
- Other chest and thoracic procedures

International Classification of Diseases, 9th Revision, Clinical Modification (ICD-9) codes as well as International Classification of Disease, 10th Revision (ICD-10) codes used to define the cohort are listed in the attached Data Dictionary.

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

The CABG surgery mortality measure excludes index admissions for patients:

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

2. Discharged against medical advice (AMA).

For patients with more than one qualifying CABG surgery admission in the measurement period, the first CABG admission is selected for inclusion in the measure and the subsequent CABG admission(s) are excluded from the cohort.

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.)*

The CABG surgery mortality measure excludes index admissions for patients:

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

Rationale: We do not include stays for patients where the age (indicated in the claim) is greater than 115, where the gender (indicated in the claim) is neither male nor female, where the admission date (indicated in the claim) is after the date of death in the Medicare Enrollment Database, or where the date of death (in the Medicare Enrollment Database) occurs before the date of discharge but the patient was discharged alive (indicated in the claim).

2. Discharged against medical advice (AMA).

Rationale: Providers did not have the opportunity to deliver full care and prepare the patient for discharge. This information is taken from the discharge disposition in the claim.

3. With more than one qualifying CABG surgery admission in the measurement period.

Rationale: CABG procedures are expected to last for several years without the need for revision or repeat revascularization. A repeat CABG procedure during the measurement period likely represents a complication of the original CABG procedure and is a clinically more complex and higher risk surgery. Therefore, we select the first CABG surgery admission for inclusion in the measure and exclude subsequent CABG surgery admissions (additional claims indicating a CABG procedure was performed within 30-days of the index CABG procedure) from the cohort.

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 for CABG surgery using a 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 the procedure date using age, sex, selected clinical covariates, and a hospital-specific effect. 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 mortality 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 (Normand and Shahian, 2007). If there were no differences among hospitals, then after adjusting for patient risk, the hospital effects should be identical across all hospitals.

The RSMR is calculated as the ratio of the number of “predicted” deaths to the number of “expected” deaths at a given hospital, multiplied by the national observed mortality rate. For each hospital, the numerator of the ratio is the number of deaths within 30 days predicted based on the hospital’s performance with its observed case mix, and the denominator is the number of deaths expected based on 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 a particular hospital’s performance, given its case mix, to be compared to an average hospital’s performance with the same case mix. Thus, a lower ratio indicates lower-than-expected mortality rates or better quality, while a higher ratio indicates higher-than-expected mortality rates 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 effect on the risk of mortality. The estimated hospital-specific effect is added to the sum of the estimated regression coefficients multiplied by the patient characteristics. The results are log 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 effect using all hospitals in our sample is added in place of the hospital-specific effect. The results are log 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 mortality rate. The hierarchical logistic regression models are described fully in the original methodology report (Suter et al. 2012).

Reference:

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

2. Suter L, Wang C, Araas M, et al. Hospital-Level 30-day All-Cause Mortality Following Coronary

Artery Bypass Graft Surgery; Updated Measure Methodology Report. 2012

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.

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

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 for the Medicare FFS measure:

Medicare Part A inpatient and Part B outpatient claims: This data source contains claims data for FFS inpatient and outpatient services including: Medicare inpatient hospital care, outpatient hospital services, as well as inpatient and outpatient physician claims for the 12 months prior to an index admission.

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 as well as vital status. These data have previously been shown to accurately reflect patient vital status (Fleming et al., 1992).

The American Community Survey (2008-2012): The American Community Survey data is collected annually and an aggregated 5-years data was used to calculate the AHRQ socioeconomic status (SES) composite index score.

Data sources for the all-payer testing: For our analyses to examine use in all-payer data, we used all-payer data from California. California is a diverse state, and, with more than 37 million residents, California represents 12% of the US population. We used the California Patient Discharge Data, a large linked database of patient hospital admissions. In 2006, there were approximately 3 million adult discharges from more than 450 non-Federal acute care hospitals. Records are linked by a unique patient identification number, allowing us to determine patient history from previous hospitalizations and to evaluate rates of both readmission and mortality (via linking with California vital statistics records).

Using all-payer data from California, we performed analyses to determine whether the HF readmission measure can be applied to all adult patients, including not only FFS Medicare patients aged 65 years or older, but also non-FFS Medicare patients aged 18-64 years at the time of admission.

Reference:

Fleming C., Fisher ES, Chang CH, Bubolz D, Malenda J. Studying outcomes and hospital utilization in the elderly: The advantages of a merged data base for Medicare and Veterans Affairs Hospitals. Medical Care. 1992; 30(5): 377-91.

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

NQF_2558_CABG_Mortality_Data_Dictionary_v1.0_final.xlsx,nqf_testing_attachment_7.1_CABG_mortality_v1.0_final-636522389533120623.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.

No

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*)**:** 2558

Measure Title: Hospital 30-Day, All-Cause, Risk-Standardized Mortality Rate (RSMR) Following Coronary Artery Bypass Graft (CABG) Surgery
Date of Submission: 1/5/2018

Type of Measure:

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

Instructions

- Measures must be tested for all the data sources and levels of analyses that are specified. ***If there is more than one set of data specifications or more than one level of analysis, contact NQF staff*** about how to present all the testing information in one form.
- **For all measures, sections 1, 2a2, 2b1, 2b2, and 2b4 must be completed.**
- **For outcome and resource use measures**, section **2b3** also must be completed.
- If specified for **multiple data sources/sets of specificaitons** (e.g., claims and EHRs), section **2b5** also must be completed.
- Respond to all questions as instructed with answers immediately following the question. All information on testing to demonstrate meeting the sub-criteria 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 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:	☒ other: Census Data/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 datasets used for testing included Medicare Parts A and B claims as well as the Medicare Enrollment Database (EDB). In addition, we used clinical data from New York State Cardiac Surgery Reporting System, and the California all-payer dataset. To assess socioeconomic factors, we used census as well as claims data (dual eligible status 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?   2008-2016

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*)

For this measure, hospitals are the measured entities. All non-federal, acute inpatient US hospitals (including territories) with Medicare Fee-for-Service (FFS) beneficiaries aged 65 years or over are included. The number of measured entities (hospitals) varies by testing type; see Section 1.7 for details.

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

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

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

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

For Reliability Testing

For reliability testing, we randomly split **Dataset 1** into two samples. The reliability of the model was tested by randomly selecting 50% of the Medicare patients aged 65 years and over in the most recent three-year cohort and calculating the risk-standardized mortality rates for this group. We then calculated risk-standardized mortality rates in the remaining 50% of patients and compared the results from each sample to assess reliability of the measure score (**Dataset 1** below).

Dataset 1 (2017 public reporting cohort): Medicare Part A Inpatient and Outpatient and Part B Outpatient claims Data

Dates of Data: July 1, 2013 – June 30, 2016

Number of Admissions: 138,661

Patient Descriptive Characteristics: average age=73.7, % male=71.7

Number of Measured Hospitals: 1,185

For Validity Testing

We assessed the face validity of the measure score using the Technical Expert Panel (TEP).

For Testing of Measure Exclusion

Dataset 1 (2017 public reporting cohort)

For Testing of Measure Risk Adjustment

Dataset 1 (2017 public reporting cohort)

Dataset 2 (development dataset): Medicare Part A Inpatient and Outpatient and Part B Outpatient claims Data

Dates of Data: January 1, 2008 – December 31, 2010

Number of Admissions: 173,291

Patient Descriptive Characteristics: average age=81.0, % male=46.1

Number of Measured Hospitals: 1,170

Dataset 3 (clinical data): New York State Cardiac Surgery Reporting System (CSRS) - New York Department of Health

Dates of Data: July 1, 2008 – June 30, 2010

Number of Admissions: 8,228

Patient Descriptive Characteristics: average age= --, % male=67.8

Number of Measured Hospitals: 35

For Optional Additional Testing for Risk Adjustment (in an adult all-payer popultaion of patients who were 18 years and older)

Dataset 4 (all-payer dataset): California all-payer dataset 2006

Dates of Data: July 1, 2013 – June 30, 2016

Number of Admissions: 14,889

Patient Descriptive Characteristics: average age=66, % male=74.9

Number of Measured Hospitals: About 450

For Testing of Measure Exclusions

Dataset 1 (2017 public reporting cohort): Medicare Part A Inpatient and Outpatient and Part B Outpatient claims Data

For Testing to Identify Meaningful Differences in Performance

Dataset 1 (2017 public reporting cohort): Medicare Part A Inpatient and Outpatient and Part B Outpatient claims Data

For Testing of Social Risk Factors in Risk Models

Dataset 1 (2017 public reporting cohort): Medicare Part A Inpatient and Outpatient and Part B Outpatient claims Data

Dataset 5: The American Community Survey (2008-2012)

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 derived from the American Community Survey (2008-2012) (**Dataset 2**) to study the association between performance measures and SES.

Data Elements Tested

- Dual eligible status (i.e., enrolled in both Medicare and Medicaid) patient-level data is obtained from CMS enrollment data (**Dataset 1**)
- Validated AHRQ SES index score is a composite of 7 different variables found in the census data

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 SES variables to analyze after reviewing the literature and examining available national data sources. There is a large body of literature linking various SES factors to worse health status and higher mortality over a lifetime (see, e.g., van Oeffelen et al., 2012). Income, education, and occupation are the most commonly examined socioeconomic factors studied. However, the literature contains few studies directly examining how different SES factors might influence the likelihood of older, insured, Medicare patients dying within 30 days of an admission for a CABG procedure. The causal pathways for SES variable selection are described below in Section 2b4.3.

The SES variables used for analysis were:

- Dual eligible status (Dataset 1)
- AHRQ-validated SES index score using 9-digit zip code data (percentage of people in the labor force who are unemployed, percentage of people living below poverty level, median household income, median value of owner-occupied dwellings, percentage of people ≥25 years of age with less than a 12th-grade education, percentage of people ≥25 years of age completing ≥4 years of college, and percentage of households that average ≥1 people per room) (Dataset 5).

References

Boan AD, Feng WW, Ovbiagele B, et al. Persistent racial disparity in stroke hospitalization and economic impact in young adults in the buckle of stroke belt. Stroke; a journal of cerebral circulation. Jul 2014;45(7):1932-1938.

Clark CJ, Guo H, Lunos S, et al. Neighborhood cohesion is associated with reduced risk of stroke mortality. Stroke; a journal of cerebral circulation. May 2011;42(5):1212-1217.

Glymour MM, Kosheleva A, Boden-Albala B. Birth and adult residence in the Stroke Belt independently predict stroke mortality. Neurology. Dec 1 2009;73(22):1858-1865.

Howard VJ, Kleindorfer DO, Judd SE, et al. Disparities in stroke incidence contributing to disparities in stroke mortality. Ann Neurol 2011;69:619–627.

Khan JA, Casper M, Asimos AW, et al. Geographic and sociodemographic disparities in drive times to Joint Commission-certified primary stroke centers in North Carolina, South Carolina, and Georgia. Preventing chronic disease. Jul 2011;8(4):A79.

Pedigo A, Seaver W, Odoi A. Identifying unique neighborhood characteristics to guide health planning for stroke and heart attack: fuzzy cluster and discriminant analyses approaches. PloS one. 2011;6(7):e22693.

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*)

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 naturally 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 compare 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 the 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) (Shrout and Fleiss, 1979), and assessed the values according to conventional standards (Landis and Koch, 1977). Specifically, we used Dataset 1 split sample and calculated the RSMR for each hospital for each sample. The agreement of the two RSMRs 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, in comparison to 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 intrarater, interrater and test–retest reliability of continuous measurements. Statistics in Medicine 2002;21:3431-3446.

Shrout P, Fleiss J. Intraclass 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*)

Measure Score Reliability Results (Dataset 1)

There were 138,661 admissions in the 2017 public reported CABG mortality measure (**Dataset 1**), with 69,040 in one sample and 69,621 in the other randomly selected sample. The agreement between the two RSMRs for each hospital was 0.35, which according to the conventional interpretation is "fair" (Landis J & Koch G, 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 the mean reliability score to be 0.851. This is considered to be high (Yu, Mehrotra, and Adams, 2013).

References

Adams J, 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.

Simoes J, Grady J, DeBuhr J et al., 2017 Procedure-Specific Measure Updates and Specifications Report Hospital-Level 30-Day Risk-Standardized Mortality Measure Isolated Coronary Artery Bypass Graft (CABG) Surgery – Version 4.0. Available at:

https://www.qualitynet.org/dcs/ContentServer?c=Page&pagename=QnetPublic%2FPage%2FQnetTier2&cid=1163010398556

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

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 ICC demonstrates fair agreement in measure score reliability.

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 risk-adjusted outcome measures, 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 (Cruz et al.). Each patient was queried twice, once by a clinician and once by a trained 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).

Our test-retest reliability score of 0.35 represents the lower bound of any reliability estimate. Using the approach used by Adams et al (2010), we obtained mean reliability score of 0.851. This pattern was also observed by Yu, Mehrotra and Adams (2013). For example, they found mean reliability for a PCP visits utilization measure to be 0.94 using the approach used by Adams and colleagues (2010), although the rest-retest reliability score was 0.68. Taking together these results indicate 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)*

Measure validity is demonstrated through prior validity testing done on our other claims-based measures, through use of established measure development guidelines, and by systematic assessment of measure face validity by a TEP of national experts and stakeholder organizations.

Face Validity as Determined by TEP
To systematically assess face validity, we surveyed the TEP and asked each member to rate the following statement using a six-point scale (1=Strongly Disagree, 2=Moderately Disagree, 3=Somewhat Disagree, 4=Somewhat Agree, 5= Moderately Agree, and 6=Strongly Agree): "The mortality rates obtained from the mortality measure as specified will provide an accurate reflection of quality."
Measure Score Validity - Validity as Assessed by External Groups

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

The working group was comprised of two cardiothoracic surgeons with expertise in quality measure development, one of whom was the lead for the development of the Society of Thoracic Surgeons (STS) registry-based CABG readmission measure. In addition, two members of the claims-based measure development team served on the working group for the STS CABG readmission measure. Through frequent (weekly or more frequent) conference calls, all aspects of measure development were discussed among the two measure developers, including the cohort definitions, outcome attribution, and risk-adjustment. The collaboration allowed real-time harmonization of the measures throughout the entire measure development process. The working group meetings addressed key issues surrounding measure development, including detailed discussions regarding the appropriate cohort for inclusion in the measure. The working group provided a forum for focused expert review and discussion of technical issues during measure development prior to consideration by the broader, combined TEP, which was convened to address all three CABG outcome measures under development (the two claims-based readmission and mortality measures as well as the registry-based readmission measure). This allowed for continuation of the close collaboration between measure developers achieved earlier in measure development.

In addition to the working group, and in alignment with the CMS Measure Management System (MMS), we convened a TEP to provide input and feedback during measure development from a group of recognized experts in relevant fields. To convene the TEP, we released a public call for nominations and selected individuals to represent a range of perspectives including clinicians, consumers, and purchasers, as well as individuals with experience in quality improvement, performance measurement, and health care disparities. We held three structured TEP conference calls consisting of presentation of key issues, our proposed approach, and relevant data, followed by open discussion among TEP members. We made minor modifications to the measure cohort (i.e., excluding additional concomitant non-cardiac procedures from the cohort such as lung resection and mastectomy), and risk-adjustment variables (i.e., including a history of prior CABG surgery in the risk adjustment) based on TEP feedback on the measures.

Following completion of the model, we solicited public comment on the measure through the CMS site link https://www.CMS.gov/MMS/17_CallforPublicComment.asp. The public comments were then posted publicly for 30 days.

Data Element Validity – Validity of Claims-Based Measures

Our team has demonstrated for a number of prior measures the validity of claims-based measures for profiling hospitals by comparing either the measure results or individual data elements against medical records. CMS validated the six

NQF-endorsed, claim-based measures currently in public reporting (acute myocardial infarction (AMI), heart failure, and pneumonia mortality and readmission) with models that used medical record-abstracted data for risk adjustment. Specifically, claims model validation was conducted by building comparable models using abstracted medical record data for risk adjustment for heart failure patients (National Heart Failure data), AMI patients (Cooperative Cardiovascular Project data) and pneumonia patients (National Pneumonia Project dataset). When both models were applied to the same patient population, the hospital risk-standardized rates estimated using the claims-based risk-adjustment models had a high level of agreement with the results based on the medical record model, thus supporting the use of the claims-based models for public reporting. Our group has reported these findings in the peer-reviewed literature (Krumholz et al. 2006; Krumholz et al. 2011; Krumholz et al. 2006a; Keenan et al. 2008; Bratzler 2011; Lindenauer 2011).

Measure Score Validity -Validity Indicated by Established Measure Development Guidelines

We developed this measure in consultation with national guidelines for publicly reported outcome measures, with outside experts, and with the public. The measure is consistent with the technical approach to outcome measurement set forth in NQF guidance for outcome measures, CMS 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; NQF 2010).

Validation of the Administrative Risk-Adjustment Model

To validate the administrative risk-adjustment model, we calculated hospital-level, RSMRs using the claims-based CABG mortality measure risk model and a risk model created using clinical registry data in a common cohort of isolated CABG patients (2008-2010) from the New York State Cardiac Surgery Reporting System (CSRS) from the New York Department of Health and compared the results. We matched claims from the 2008-2010 data sets to the 2008-2011 NY Registry data.

We measured the correlation between the two sets of results at the hospital level. In addition, we used a bootstrapping approach similar to that used for public reporting of the AMI, heart failure and pneumonia mortality measures to categorize hospital performance as better, worse or no different than the average hospital observed mortality rate. The bootstrapping algorithm used is described below:

Let *I* denote the total number of hospitals in the sample. We repeat steps 1 – 4 below for b = 1,2,…B times:

1. Sample *I* hospitals with replacement.
2. Fit the hierarchical logistic regression model defined by Equation (1) using all patients within each sampled hospital. The starting values are the parameter estimates obtained by fitting the model to all hospitals. If some hospitals are selected more than once in a bootstrapped sample, we treat them as distinct so that we have *I* random effects to estimate the variance components. After Step 2, we have:
    a. The estimated regression coefficients of the risk factors, $\widehat{\boldsymbol{\beta}}^{(b)}$.
    b. The parameters governing the random effects, hospital adjusted outcomes, distribution $\hat{\mu}^{(b)}$ and $\hat{\tau}^{2(b)}$.
    c. The set of hospital-specific intercepts and corresponding variances,$\left\{\hat{\alpha}_i^{(b)}, v\hat{a}r\left(\alpha_i^{(b)}\right); i = 1,2,\ldots,I\right\}$
3. We generate a hospital random effect by sampling from the distribution of the hospital-specific distribution obtained in Step 2c. We approximate the distribution for each random effect by a normal distribution. Thus, we draw $\alpha_i^{(b*)} \sim N(\hat{\alpha}_i^{(b)}, v\hat{a}r\left(\alpha_i^{(b)}\right))$ for the unique set of hospitals sampled in Step 1.
4. Within each unique hospital *i* sampled in Step 1, and for each case *j* in that hospital, we calculate $\hat{p}_{ij}^{(b)}$, $\hat{e}_{ij}^{(b)}$, and $\hat{s}_i^{(b)}$ where $\widehat{\boldsymbol{\beta}}^{(b)}$ and $\hat{\mu}^{(b)}$ are obtained from Step 2 and $\alpha_i^{(b*)}$ is obtained from Step 3.

    Ninety-five percent interval estimates (or alternative interval estimates) for the hospital-standardized outcome can be computed by identifying the 2.5th and 97.5th percentiles of the B estimates (or the percentiles corresponding to the alternative desired intervals).

We then performed a reclassification analysis to determine how many hospitals might be reclassified to a different performance category if assessed by the administrative model as compared to the registry model. In order to isolate differences due to the method of risk adjustment, both measures were calculated in the same cohort of patients, used the same outcome definition (30-day all-cause mortality defined by administrative claims data) and a consistent approach to risk-adjustment modeling (the hierarchical logistic regression model approach used in CMS's publicly reported claims-based outcome measures).

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

We re-specified the measure 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.)
- We re-specified the risk model:
- The CC-based risk variables were updated to the ICD-10-compatible HCC system version 22, maintained by RTI International; and,
- Certain risk variables (for example, cardiogenic shock), previously defined using ICD-9 codes, were re-defined using ICD-10 codes for use with inpatient, outpatient, and/or physician Medicare administrative claims on or after October 1, 2015.

The goal of this re-specification was to maintain the intent and validity of the measure.

The ICD-10 Transition Process

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. 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 a set 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.

The risk variables were updated to the ICD-10-compatible HCC version 22 map. The intent was to keep the risk-adjustment model as similar as possible to the model previously defined using HCC version 12. Specifically:

- Experts examined the ICD-9 code-based HCC version 12 and version 22 maps and reviewed shifts that occurred (where an ICD-9 code had moved from one CC to another). Based on these examinations, they recommended new risk variables using version 22 CCs.
- Following re-specification of the risk variables using the HCC version 22 map, we ran risk-adjustment models on several outcome measures, to ensure testing of all variables where shifts in the ICD-9 codes included in the CCs had occurred.
- For each tested measure, we used the same claims dataset to calculate and compare two separate sets of measure results using two separate risk-adjustment models: One set using the previously-specified version 12 risk variables, and the other using the newly-specified version 22 risk variables. For this analysis, we used the ICD-9-coded data from the 2016 measurement period.
- We compared the frequencies and model coefficients of the two sets of risk-adjustment variables, to ensure that they were similar.
- We compared the performance of each risk-adjustment model by calculating each model's c-statistic and predictive ability.
- We examined the correlation in the risk-standardized outcome rates produced by the two risk-adjustment models, to ensure that they produced similar measure results.
- We examined the degree to which the models produced similar risk-standardized outcome rates at the hospital level by assessing whether individual hospitals' risk-standardized rates fell into the same quintile in the distribution of risk-standardized rates calculated by each of the two models.
- Based on the results of these analyses, we made minor modifications to the re-specified risk-adjustment variables to ensure that the performance of the risk-adjustment model was as similar as possible to the performance of the previously-specified model, and that the hospital-level results were as similar as possible.

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

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.

Xian Y, Fonarow GC, Reeves MJ, et al. Data quality in the American Heart Association Get With The Guidelines-Stroke (GWTG-Stroke): Results from a National Data Validation Audit. American Heart Journal. 2012;163(3):392-398.e391. http://www.ahjonline.com/article/S0002-8703%2811%2900894-5/abstract

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

Validity as Assessed by External Groups

Fourteen TEP members responded to the survey question as follows: Strongly Disagreed (1), Moderately Disagreed (1), Somewhat Disagreed (1), Somewhat Agreed (1), Moderately Agreed (8), and Strongly Agreed (2). Hence, 79% of TEP members agreed (71% moderately or strongly agreed) that the measure will provide an accurate reflection of quality.

Validation of Administrative Risk Adjustment Model

The validation of the administrative risk model demonstrated similar distributions in hospital RSMRs for the claims-based and clinical-based models, although the claims-based model showed a narrower range of outcome rates. The C-statistics for the two models were similar: 0.74 for the claims-based model and 0.75 for the clinical-based model. Overall agreement between hospital performance categorization between the claims-based and clinical-based models was 94.3% (33 of 35 hospitals had concordant performance categorization) and the correlation was 0.90 (weighted Spearman correlation). The clinical-based model identified two worse-performing outlier hospitals, while the claims-based model identified none; neither model identified any better-performing outliers in the matched sample.

Full results of the validation study can be found in the Appendix of the attached CABG Mortality Measure Methodology Report.

References

Krumholz HM, Lin Z, Drye EE, Desai MM, Han LF, Rapp MT, Mattera JA, Normand SL. An administrative claims measure suitable for profiling hospital performance based on 30-day all-cause readmission rates among patients with acute myocardial infarction. Circulation: Cardiovascular Quality and Outcomes. 2011 Mar 1;4(2):243-52.

Keenan PS, Normand SL, Lin Z, Drye EE, Bhat KR, Ross JS, Schuur JD, Stauffer BD, Bernheim SM, Epstein AJ, Wang Y-F, Herrin J, Chen J, Federer JJ, Mattera JA, Wang Y, Krumholz HM. 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: Cardiovascular Quality and Outcomes. 2008 Sep;1(1):29-37.

Lindenauer PK, Normand SL, Drye EE, Lin Z, Goodrich K, Desai MM, Bratzler DW, O´Donnell WJ, Metersky ML, Krumholz HM. Development, validation, and results of a measure of 30-day readmission following hospitalization for pneumonia. Journal of Hospital Medicine. 2011 Mar;6(3):142-50.

Shahian DM, Silverstein T, Lovett AF, Wolf RE, Normand SLT. "Comparison of Clinical and Administrative Data Sources for Hospital Coronary Artery Bypass Graft Surgery Report Cards." Circulation. 2007; 115: 1518-1527.

Curtis J, Drye E, Geary L, et al. Hospital 30-Day Percutaneous Coronary Intervention Mortality Measure: Center for Outcomes Research and Evaluation;2010.

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?*)

Validity as Assessed by External Groups

The results demonstrate TEP agreement with overall face validity of the measure as specified. Measure validity is also ensured through the processes employed during development, including regular expert and clinical input, and modeling methodologies with demonstrated validity in claims-based measures.

Validation of Administrative Risk Adjustment Model

Thorough evaluation adherent to nationally accepted standards for outcome measure development (Krumholz et al. 2008; Shahian et al. 2007) indicate that the model has similar discrimination and calibration to a New York state-derived clinical risk model, although the relative discrimination was lower when a risk variable (shock), whose pre-operative status was unknown, was removed from the claims-based model. Although both the mortality rate and range of performance in the matched sample was less than that of US hospitals overall, the frequency and effect of risk variables was similar in the matched sample and national data. The models produce similar estimates of hospital performance. However, the claims-based model generally produced lower RSMR estimates compared with the clinical-based model among hospitals with higher estimated RSMRs, and higher RSMR estimates among those hospitals with lower RSMRs. Assuming that the clinical-based model is the gold standard (and does not over-estimate poor performing hospitals' RSMRs), our findings suggest that the claims-based model may underestimate poor performing hospitals' RSMRs and may be less likely to identify poor performance outliers compared with the clinical-based model. Similarly, the claims-based model may be less likely to identify hospitals with significantly better-than-average performance, although this validation study cannot assess this as the clinical-based model did not identify high performing outlier hospitals in the validation sample.

References

Shahian DM, Silverstein T, Lovett AF, Wolf RE, Normand SLT. "Comparison of Clinical and Administrative Data Sources for Hospital Coronary Artery Bypass Graft Surgery Report Cards." *Circulation.* 2007; 115: 1518-1527.

Krumholz HM, Keenan PS, Brush JE, Jr., et al. Standards for measures used for public reporting of efficiency in health care: a scientific statement from the American Heart Association Interdisciplinary Council on Quality of Care and Outcomes Research and the American College of Cardiology Foundation. *Circulation.* Oct 28 2008;118(18):1885-1893.

____

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 field S.9 of the measure submission form (Denominator Exclusions Details).

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	%	Distribution across hospitals (N=1,039): Min, 25th, 50th, 75th percentile, max
1. Inconsistent or unknown vital status or other unreliable demographic data	**1**	**<0.01%**	**(0, 0, 0, 0, 0.25)**
2. Admissions for subsequent qualifying CABG procedures during the measurement period	**88**	**0.06%**	**(0, 0, 0, 1.19, 3.45)**
3. Discharged against medical advice (AMA)	**46**	**0.03%**	**(0, 0, 0, 0, 3.70)**

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 is necessary for valid calculation of the measure. Patients with an inconsistent or unknown vital status or other unreliable demographic account for <0.01% of all index admissions excluded from the initial index cohort.

Exclusion 2 (admissions for subsequent qualifying CABG procedures during the measurement period) accounts for 0.06% of all index admissions excluded from the initial index cohort. This exclusion was applied to align with the Society of Thoracic Surgeons 30-day mortality measure. The experts believed that a second CABG procedure within 30 days of an initial procedure is most likely due to a complication of the initial CABG procedure or the peri-operative care the patient received, and as such, the care provided by the hospital performing the initial CABG procedure likely dominates mortality risk.

Exclusion 3 (patients who are discharged AMA) accounts for 0.03% 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.

____

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 24 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.

See risk model specification in Section 2b3.4a 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 goal was to develop a parsimonious model that included clinically relevant variables associated with isolated CABG mortality. The candidate variables for the model were derived from: the index admission, with comorbidities identified from the index admission secondary diagnoses (excluding potential complications), 12-month pre-index inpatient Part A data, outpatient hospital data, and Part B physician data.

For administrative model development, we started with 189 Condition Categories (CCs) which are part of CMS's Hierarchical Condition Categories. The Hierarchical Condition Category (HCC) system groups the ICD-9-CM codes into larger groups that are used in models to predict medical care utilization, mortality, or other related measures. CCs are clinically relevant diagnostic groups of the more than 15,000 ICD-9 codes (Pope et al. 2001).

To select candidate variables, a team of clinicians reviewed all 189 CCs and excluded those that were not relevant to the Medicare population or that were not clinically relevant to the mortality outcome (e.g., attention deficit disorder, female infertility). Clinically relevant CCs were selected as candidate variables and some of those CCs were then combined into clinically coherent CC groupings. Other candidate variables included age, gender, and cardiogenic shock. Gender was included in risk adjustment due to the fact that women have smaller caliber vessels and thus represent more technically challenging CABG procedures compared to men (O'Connor 1996).

To inform final variable selection, a modified approach to stepwise logistic regression was performed. The development sample was used to create 1,000 "bootstrap" samples. For each sample, we ran a logistic stepwise regression that included the candidate variables. The results were summarized to show the percentage of times that each of the candidate variables was significantly associated with mortality (p<0.001) in each of the 1,000 repeated samples (e.g., 90 percent would mean that the candidate variable was selected as significant at p<0.001 in 90 percent of the estimations). We also assessed the direction and magnitude of the regression coefficients.

The clinical team reviewed these results and decided to retain the majority of risk adjustment variables above a 70% cutoff, because they demonstrated a relatively strong and stable association with risk for death and were clinically relevant. Additionally, specific variables with particular clinical relevance to the risk of death were forced into the model (regardless of percent selection) to ensure appropriate risk adjustment for CABG. These included:

1) Clinical variables associated with CABG:
- History of Prior CABG or Valve Surgery

2) Markers for end of life/frailty:
- Decubitus Ulcer or Chronic Skin Ulcer
- Dementia or Other Specified Brain Disorders
- Metastatic Cancer and Acute Leukemia
- Protein-calorie Malnutrition
- Hemiplegia, Paraplegia, Paralysis, Functional disability
- Stroke

3) Diagnoses with potential asymmetry among hospitals that would impact the validity of the model:
- Lung, Upper Digestive Tract, and Other Severe Cancers
- Lymphatic, Head and Neck, Brain, and Other Major Cancers; Breast, Prostate, Colorectal and Other Cancers and Tumors; Other Respiratory and heart Neoplasms
- Other Digestive and Urinary Neoplasms

This resulted in a final risk-adjustment model that included 24 variables.

References

Pope G, Ellis R, Ash A, et al. Principal Inpatinet Diagnosit Cost Group Models for Medicare Risk Adjustment. *Health Care Financing Review.* 2000;21(3):26.

O'Connor NJ, Morton JR, Birkmeyer JD, Olmstead EM, O'Connor GT. Effect of coronary artery diameter in patients undergoing coronary bypass surgery. Northern New England Cardiovascular Disease Study Group. *Circulation.* 1996;93(4):652-655.

Social Risk Factors

We selected variables representing social risk factors such as SES for examination based on a review of literature, conceptual pathways, and feasibility. In Section 1.8, we describe the variables that we considered and analyzed based on this review. Below we describe the pathways by which social risk factors may influence 30-day mortality.

Our conceptualization of the pathways by which patient social risk factors affects 30-day mortality is informed by the literature.

Literature Review of Social Risk Variables and Mortality after a CABG Procedure

To examine the relationship between social risk factors and hospital 30-day, all-cause, RSMR following CABG surgery, a literature search was performed with the following exclusion criteria: international studies, articles published more than 10 years ago, articles without primary data, articles using Veterans Affairs databases as the primary data source, and articles not explicitly focused on social risk factors such as SES and CABG mortality. Studies are limited, and those that have been conducted have mixed results.

Causal Pathways for Social Risk Variable Selection

Although some recent literature evaluates the relationship between patient social risk factor such as SES and the mortality outcome, few studies directly address causal pathways or examine the role of the hospital in these pathways (see, for example, Chang et al 2007; Gopaldas et al 2009; Kim et al 2007; LaPar 2010; 2012). Moreover, the current literature examines a wide range of conditions and risk variables with no clear consensus on which risk factors demonstrate the strongest relationship with mortality. The social risk factors that have been examined in the literature can be categorized into three domains: (1) patient-level variables, (2) neighborhood/community-level variables, and (3) hospital-level variables. Patient-level variables describe characteristics of individual patients, and include the patient's income or education level (Eapen et al., 2015). Neighborhood/community-level variables use information from sources such as the American Community Survey as either a proxy for individual patient-level data or to measure environmental factors. Studies using these variables use one dimensional measures such as median household income or composite measures such as the AHRQ-validated SES index score (Blum et al., 2014). Hospital-level variables measure attributes of the hospital which may be related to patient risk. Examples of hospital-level variables used in studies are ZIP code characteristics aggregated to the hospital level or the proportion of Medicaid patients served in the hospital (Gilman et al., 2014; Joynt and Jha, 2013).

The conceptual relationship, or potential causal pathways by which these possible social risk factors influence the risk of mortality following an acute illness or major surgery, like the factors themselves, are varied and complex. There are at least four potential pathways that are important to consider.

1. **Relationship of social risk factors such as SES to health at admission**. Patients who have lower income/education/literacy or unstable housing may have a worse general health status and may present for their hospitalization or procedure with a greater severity of underlying illness. These social risk factors, which are characterized by patient-level or neighborhood/community-level (as proxy for patient-level) variables, may contribute to worse health status at admission due to competing priorities (restrictions based on job, lack of childcare), lack of access to care (geographic, cultural, or financial), or lack of health insurance. Given that these risk factors all lead to worse general health status, this causal pathway should be largely accounted for by current clinical risk-adjustment**.**

2. **Use of low-quality hospitals**. Patients of lower income, lower education, or unstable housing have been shown not to have equitable access to high quality facilities because such facilities are less likely to be found in geographic areas with large populations of poor patients. Thus, patients with low income are more likely to be seen in lower quality hospitals, which can explain increased risk of mortality following hospitalization.

3. **Differential care within a hospital**. The third major pathway by which social risk factors may contribute to mortality risk is that patients may not receive equivalent care within a facility. For example, patients with social risk factors such as lower education may require differentiated care (e.g. provision of lower literacy information – that they do not receive).

4. **Influence of social risk factors on mortality risk outside of hospital quality and health status**. Some social risk factors, such as income or wealth, may affect the likelihood of mortality without directly affecting health status at

admission or the quality of care received during the hospital stay. For instance, while a hospital may make appropriate care decisions and provide tailored care and education, a lower-income patient may have a worse outcome post-discharge due to competing economic priorities or a lack of access to care outside of the hospital.

These proposed pathways are complex to distinguish analytically. They also have different implications on the decision to risk adjust or not. We, therefore, first assessed if there was evidence of a meaningful effect on the risk model to warrant efforts to distinguish among these pathways.

Based on this model and the considerations outlined in Section 1.8, the following social risk variables were considered:

- Dual eligible status
- AHRQ SES index

We assessed the relationship between the SES variables with the outcome and examined the incremental effect in a multivariable model. For this measure, we also examined the extent to which the addition of any one of these variables improved model performance or changed hospital results. Given no meaningful improvement in the risk-model or change in performance scores we did not further seek to distinguish the causal pathways for these measures.

References

Blum, A. B., N. N. Egorova, E. A. Sosunov, A. C. Gelijns, E. DuPree, A. J. Moskowitz, A. D. Federman, D. D. Ascheim and S. Keyhani. "Impact of Socioeconomic Status Measures on Hospital Profiling in New York City." *Circ Cardiovasc Qual Outcomes* 7, no. 3 (2014): 391-7.

Calvillo-King L, Arnold D, Eubank KJ, et al. Impact of social factors on risk of readmission or mortality in pneumonia and heart failure: systematic review. *Journal of general internal medicine.* 2013;28(2):269-282.

Chang W-C, Kaul P, Westerhout C M, Graham M. M., Armstrong Paul W., “Effects of Socioeconomic Status on Mortality after Acute Myocardial Infarction.” The American Journal of Medicine. 2007; 120(1): 33-39

Eapen ZJ, McCoy LA, Fonarow GC, Yancy CW, Miranda ML, Peterson ED, Califf RM, Hernandez AF. Utility of socioeconomic status in predicting 30-day outcomes after heart failure hospitalization. Circ Heart Fail. May 2015; 8(3):473-80.

Foraker, R. E., K. M. Rose, C. M. Suchindran, P. P. Chang, A. M. McNeill and W. D. Rosamond. "Socioeconomic Status, Medicaid Coverage, Clinical Comorbidity, and Rehospitalization or Death after an Incident Heart Failure Hospitalization: Atherosclerosis Risk in Communities Cohort (1987 to 2004)." *Circ Heart Fail* 4, no. 3 (2011): 308-16.

Gopaldas R R, Chu D., “Predictors of surgical mortality and discharge status after coronary artery bypass grafting in patients 80 years and older.” The American Journal of Surgery. 2009; 198(5): 633-638

Gilman M, Adams EK, Hockenberry JM, Wilson IB, Milstein AS, Becker ER. California safety-net hospitals likely to be penalized by ACA value, readmission, and meaningful-use programs. Health Affairs (Millwood). Aug 2014; 33(8):1314-22.

Joynt KE, Jha AK. Characteristics of hospitals receiving penalties under the Hospital Readmissions Reduction Program. JAMA. Jan 23 2013; 309(4):342-3.

Kind, A. J., S. Jencks, J. Brock, M. Yu, C. Bartels, W. Ehlenbach, C. Greenberg and M. Smith. "Neighborhood Socioeconomic Disadvantage and 30-Day Rehospitalization: A Retrospective Cohort Study." Ann Intern Med 161, no. 11 (2014): 765-74.

Kim C, Diez A V, Diez Roux T, Hofer P, Nallamothu B K, Bernstein S J, Rogers M, “Area socioeconomic status and mortality after coronary artery bypass graft surgery: The role of hospital volume.” Clinical Investigation Outcomes, Health Policy, and Managed Care. 2007; 154(2): 385-390

Krumholz HM, Brindis RG, Brush JE, 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: 456-462.

LaPar D J, Bhamidipati C M, et al. “Primary Payer Status Affects Mortality for Major Surgical Operations.” Annals of Surgery. 2010; 252(3): 544-551

LaPar D J, Stukenborg G J, et al “Primary Payer Status Is Associated With Mortality and Resource Utilization for Coronary Artery Bypass Grafting.” Circulation. 2012; 126:132-139

Normand S-LT, Shahian DM. 2007. Statistical and Clinical Aspects of Hospital Outcomes Profiling. Stat Sci 22 (2): 206-226.
Reames BN, Birkmeyer NJ, Dimick JB, Ghaferi AA. Socioeconomic disparities in mortality after cancer surgery: failure to rescue. JAMA surgery 2014; 149:475-81.

Regalbuto R, Maurer MS, Chapel D, Mendez J, Shaffer JA. Joint Commission requirements for discharge instructions in patients with heart failure: is understanding important for preventing readmissions? *Journal of cardiac failure.* 2014;20(9):641-649.

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?

Below is a table showing the final variables in the model with associated odds ratios (OR) and confidence intervals (CI) (**Dataset 1**).

Variable	**07/2013-06/2016** **OR (95% CI)**
Age minus 65 (years above 65, continuous)	1.06 (1.06 - 1.07)
Male	0.69 (0.64 - 0.74)
Cardiogenic shock	7.20 (6.68 - 7.75)
Coronary atherosclerosis	1.18 (1.06 - 1.33)
History of coronary artery bypass graft (CABG) or valve surgery	1.41 (1.24 - 1.60)
Cancer; metastatic cancer and acute leukemia (CC 8-14)	0.92 (0.84 - 1.00)
Protein-calorie malnutrition (CC 21)	1.72 (1.55 - 1.91)
Morbid obesity; other endocrine/metabolic/nutritional disorders (CC 22, 25-26)	0.73 (0.66 - 0.82)
Liver or biliary disease (CC 27-32)	1.50 (1.35 - 1.67)
Other gastrointestinal disorders (CC 38)	0.77 (0.72 - 0.82)
Dementia or other specified brain disorders (CC 51-53)	1.29 (1.16 - 1.45)
Hemiplegia, paraplegia, paralysis, functional disability (CC 70-74, 103-104, 189-190)	1.29 (1.10 - 1.52)
Congestive heart failure (CC 85)	1.17 (1.08 - 1.27)
Acute myocardial infarction (CC 86)	1.20 (1.11 - 1.29)
Unstable angina and other acute ischemic heart disease (CC 87)	0.87 (0.81 - 0.93)
Angina; old myocardial infarction (CC 88 plus ICD-10-CM code I25.2, for discharges on or after October 1, 2015; CC 88 plus ICD-9-CM code 412, for discharges prior to October 1, 2015)	0.87 (0.81 - 0.93)
Hypertension (CC 95)	0.81 (0.74 - 0.89)

Variable	07/2013-06/2016 OR (95% CI)
Stroke (CC 99-100)	1.06 (0.92 - 1.22)
Vascular or circulatory disease (CC 106-109)	1.16 (1.08 - 1.24)
Chronic obstructive pulmonary disease (COPD) (CC 111)	1.38 (1.29 - 1.48)
Pneumonia (CC 114-116)	1.32 (1.21 - 1.43)
Dialysis status (CC 134)	1.92 (1.66 - 2.23)
Renal failure (CC 135-140)	1.39 (1.30 - 1.49)
Decubitus ulcer or chronic skin ulcer (CC 157-161)	1.11 (0.97 - 1.28)

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.**

Variation in prevalence of the factor across measured entitiesThe prevalence of SES factors in the CABG cohort varies across measured entities. The median percentage of dual eligible patients is 8.4% (interquartile range [IQR]: 5.6% – 13.4%). The median percentage of patients with an AHRQ SES Index score equal to or below 46.0 is 16.6% (IQR: 8.8% – 26.8%).

Empirical association with the outcome (univariate)

The patient-level observed CABG mortality rate is higher for dual eligible patients, 4.68%, compared with 3.03% for all other patients. Similarly, the mortality rate for patients with an AHRQ SES Index score equal to or below 42.6 was 3.96% compared with 3.00% for patients with an AHRQ SES Index score above 42.6.

Incremental effect of SES variables in a multivariable model

We then examined the strength and significance of the SES variables in the context of a multivariable model. Consistent with the above findings, when we include any of these variables in a multivariate model that includes all of the claims-based clinical variables, the effect size of each of these variables is significant, but lower, than the coefficient for the bivariate association (the parameter estimate decreased from 1.57 to 1.23 for dual eligibility, from 1.34 to 1.23 for the AHRQ SES Index).

To further understand the relative importance of these risk-factors in the measure, we compared hospital performance with and without the addition of each social risk variable. Results show that the c-statistic is unchanged with the addition of any of these variables into the model: The c-statistic of the original model is 0.779; the c-statistic of the original model with the dual eligible variable added is 0.779; and the original model with the AHRQ SES index variable added is 0.780.

We also examined the change in hospitals' RSMRs with the addition of any of these variables. The median absolute change in hospitals' RSMRs when adding a dual eligibility indicator is 0.010% (IQR: 0.00% – 0.03%, minimum 0.00% – maximum 0.48%) with a correlation coefficient between RSMRs for each hospital with and without dual eligibility added of 0.99781. The median absolute change in hospitals' RSMRs when adding an indicator for a low AHRQ SES Index score is 0.02% (IQR: 0.01% – 0.03%, minimum 0.00% – maximum 0.22%) with a correlation coefficient between RSRRs for each hospital with and without an indicator for a low AHRQ SES Index score added of 0.99961.

Overall, we find that the social risk variables that could be feasibly incorporated into this model do have a significant relationship with the outcome in multivariable modeling. However, the impact of any of these indicators is very small to negligible on model performance and hospital profiling. Given the controversial nature of incorporating such variables into a risk-model we do not support doing so in a case that is unlikely to affect hospital profiling.

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*)

Approach to assessing model performance (**Dataset 1** and **Dataset 2**)

We computed three summary statistics for assessing model performance (Harrell and Shih, 2001) for the CABG mortality cohort:

Discrimination Statistics

(1) Area under the receiver operating characteristic (ROC) curve (the c-statistic) is the probability that predicting the outcome is better than chance, which is a measure of how accurately a statistical model is able to distinguish between a patient with and without an outcome)

(2) Predictive ability (discrimination in predictive ability measures the ability to distinguish high-risk subjects from low-risk subjects; therefore, we would hope to see a wide range between the lowest decile and highest decile.)

Calibration Statistics

(3) Over-fitting indices (over-fitting refers to the phenomenon in which a model accurately describes the relationship between predictive variables and outcome in the development dataset but fails to provide valid predictions in new patients)

We tested the performance of the model for **Dataset 1** described in section 1.7.

References

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

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

If stratified, skip to 2b3.9

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

Results for the development cohort (Dataset 2)

2009 development cohort:
C-statistic = 0.75
Predictive ability (lowest decile %, highest decile %): (0.7, 11.1)
2008 validation cohort:
C-statistic = 0.74
Predictive ability (lowest decile %, highest decile %): (0.6, 11.8)
2010 validation cohort:
C-statistic = 0.75

Predictive ability (lowest decile %, highest decile %): (0.5, 10.6)

Results for the 2017 reporting cohort (Dataset 1)

C statistic = 0.7789;

Predictive ability (lowest decile %, highest decile %) = (0.4, 14.0)

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

Results for the development cohort (Dataset 2)

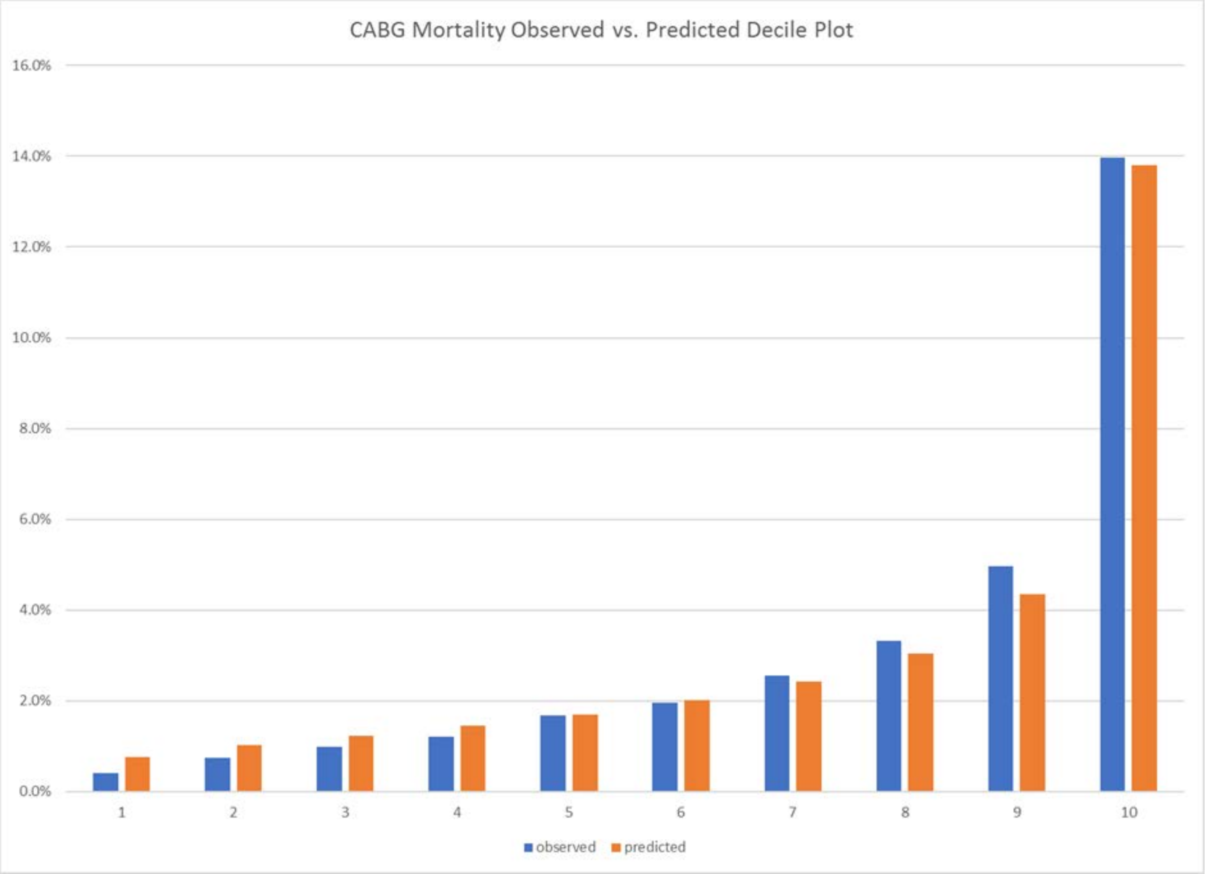
2009 development cohort: Calibration (over-fitting statistics): (0, 1)

2008 validation cohort: Calibration (over-fitting statistics): (0.01, 0.99)

2010 validation cohort: Calibration (over-fitting statistics): (-0.10, 0.97)

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

The risk decile plot is a graphical depiction of the deciles calculated to measure predictive ability. Below, we present the risk decile plot showing the distributions for Medicare FFS data from July 2013 to June 2016 (**Dataset 1**).



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 (Dataset 1 and Dataset 2)

The C-statistics ranged from 0.75 to 0.78 across datasets (Dataset 1 and Dataset 2) and indicates good model discrimination. The model indicated a wide range between the lowest decile and highest decile, indicating the ability to distinguish high-risk subjects from low-risk subjects.

Calibration Statistics (Dataset 2)

Over-fitting (Calibration γ0, γ1)

If the γ0 in the validation samples are substantially far from zero and the γ1 is substantially far from 1, there is potential evidence of over-fitting. The calibration value of close to zero at one end and close to 1 on the other end indicates good calibration of the model (Dataset 2).

Risk Decile Plots (Dataset 1)

Higher deciles of the predicted outcomes are associated with higher observed outcomes, which show a good calibration of the model. This plot indicates excellent discrimination of the model and good predictive ability.

Overall Interpretation (Dataset 1 and Dataset 2)

Interpreted together, our diagnostic results demonstrate the 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*)

Application to Patients Aged 18 Years and Older (Dataset 4)

When the model was applied to all patients aged 18+ in 2006 California Patient Discharge Data, overall discrimination was good (C statistic=0.84). In addition, there was good discrimination and predictive ability in both those aged 18-64 and those aged 65+. Moreover, the distribution of Pearson residuals was comparable across the patient subgroups. When comparing the model with and without interaction terms [between age (>65 and <65) and individual risk factors]: (a) the reclassification analysis demonstrated nearly 100% overall agreement in patient risk categorization; (b) the C statistic was nearly identical for the models with and without interaction terms (0.85 vs. 0.86, respectively); and (c) hospital-level risk-standardized rates were highly correlated (ICC=0.998). Although there were significant age-by-risk-factor interaction terms for two variables (Older and COPD, and Older and Dementia or Senility), the inclusion of interactions did not substantively affect either patient-level model performance or hospital-level results. Therefore, the measure can be applied to all-payer data for patients 18 years and older. For simplicity and pending further study, the only change currently recommended to the measure specifications to allow application to an all-payer, 18+ year population is transformation of the Age variable from "Age – 65" to a fully continuous age variable.

____

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 public reporting of the measure, 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.

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 of Medicare FFS data show substantial variation in RSMRs among hospitals. Using data from July 2013-June 2016 (**Dataset 1**), the median hospital RSMR was 3.1%, with a range of 1.3% to 7.4%. The interquartile range was 2.7%-3.7%.

Of 1,185 hospitals in the study cohort, 17 performed "Better than the National Rate," 1,004 performed "No Different from the National Rate," and 18 performed "Worse than the National Rate." 146 were classified as "Number of Cases Too Small" (fewer than 25) to reliably tell how well the hospital is performing.

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 30-day all-cause mortality following a qualifying CABG procedure.

____

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*)

N/A

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*)

N/A

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*)

N/A

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.

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 electronic claims

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).

N/A

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.

Administrative data are routinely collected as part of the billing process. Because completion of claims is required for hospital reimbursement, there is little missing data. The measures do not require any additional data collection.

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 associated with the 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)
Payment Program Not in use	Public Reporting Hospital Inpatient Quality Reporting (IQR) Program https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/HospitalQualityInits/HospitalRHQDAPU.html

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

Public Reporting

Program Name, Sponsor: Hospital Inpatient Quality Reporting (Hospital IQR) Program, Centers for Medicare and Medicaid Services (CMS)

Purpose: The Hospital Inpatient Quality Reporting Program was originally mandated by Section 501(b) of the Medicare Prescription Drug, Improvement, and Modernization Act (MMA) of 2003. This section of the MMA authorized CMS to pay hospitals that successfully report designated quality measures a higher annual update to their payment rates. Initially, the MMA provided for a 0.4 percentage point reduction in the annual market basket (the measure of inflation in costs of goods and services used by hospitals in treating Medicare patients) update for hospitals that did not successfully report. The Deficit Reduction Act of 2005 increased that reduction to 2.0 percentage points. This was modified by the American Recovery and Reinvestment Act of 2009 and the Affordable Care Act of 2010, which provided that beginning in fiscal year (FY) 2015, the reduction would be by one-quarter of such applicable annual payment rate update if all Hospital Inpatient Quality Reporting Program requirements are not met.

Under the Hospital Inpatient Quality Reporting Program, CMS collects quality data from hospitals paid under the Inpatient Prospective Payment System, with the goal of driving quality improvement through measurement and transparency by publicly displaying data to help consumers make more informed decisions about their health care. It is also intended to encourage hospitals and clinicians to improve the quality and cost of inpatient care provided to all patients. The data collected through the program are available to consumers and providers on the Hospital Compare website at: https://www.medicare.gov/hospitalcompare/search.html. Data for selected measures are also used for paying a portion of hospitals based on the quality and efficiency of care, including the Hospital Value-Based Purchasing Program, Hospital-Acquired Condition Reduction Program, and Hospital Readmissions Reduction Program.

Geographic area and number and percentage of accountable entities and patients included:

The Hospital IQR program includes all Inpatient Prospective Payment System (IPPS), non-federal, acute care hospitals and VA hospitals in the United States. The number and percentage of accountable entities included in the program, as well as the number of patients included in the measure, varies by reporting year. For the data period between 2013-

2016, the number of hospitals included in the measure with the expanded cohort was 1,185 and the number of admissions was 138,661.

Payment Program

Program Name, Sponsor: Hospital Value-Based Purchasing, Centers for Medicare and Medicaid Services (CMS)

Purpose:

The Hospital VBP Program is designed to promote better clinical outcomes for hospital patients, as well as improve their experience of care during hospital stays. Specifically, Hospital VBP seeks to encourage hospitals to improve the quality and safety of care that Medicare beneficiaries and all patients receive during acute-care inpatient stays by: Eliminating or reducing the occurrence of adverse events (healthcare errors resulting in patient harm); adopting evidence-based care standards and protocols that result in the best outcomes for most patients; re-engineering hospital processes that improve patients' experience of care; increasing the transparency of care for consumers; and recognizing hospitals that are involved in the provision of high-quality care at a lower cost to Medicare.

Geographic area and number and percentage of accountable entities and patients included:

As defined in Social Security Act Section 1886(d)(1)(B), the program applies to subsection (d) hospitals located in the 50 states and the District of Columbia. The following categories of hospitals are excluded from the program: Hospitals subject to payment reductions under the Hospital Inpatient Quality Reporting (IQR) Program; hospitals excluded from the Inpatient Prospective Payment System (IPPS), such as psychiatric, rehabilitation, long-term care, children's, critical access, and 11 Prospective Payment System (PPS)-exempt cancer hospitals; hospitals located in Puerto Rico and other United States territories are also excluded; hospitals located in the state of Maryland participating in the Maryland All-Payer Model; hospitals cited for deficiencies during the applicable fiscal year performance period(s) that pose an immediate jeopardy (IJ) to patients' health or safety; and hospitals with an approved extraordinary circumstance exemption specific to the Hospital VBP Program.

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?*)

N/A. This measure is currently publicly reported.

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.*)

N/A. This measure is currently publicly reported.

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.

The exact number of measured entities (acute care hospitals) varies with each new measurement period. In 2017, 1,185 hospitals were included in measure calculation. These were all of the U.S. Section D, and critical access hospitals with at least 25 CABG cases performed between July 2013 and June 2016.

Each hospital receives their measure results in April of each calendar year through CMS's QualityNet website. The results are then publicly reported on CMS's Hospital Compare website in July of each calendar year. Since the measure is risk standardized using data from all hospitals, hospitals cannot independently calculate their score.

However, CMS provides each hospital with several resources that aid in the interpretation of their results (described in detail below). These include Hospital-Specific Reports with details about every patient from their facility that was included in the measure calculation (for example, dates of admission and discharge, discharge diagnoses, outcome [died or not], transfer status, and facility transferred from). These reports facilitate quality improvement activities such as review of individual deaths and patterns of deaths; make visible to hospitals post-discharge outcomes that they may otherwise be unaware of; and allow hospitals to look for patterns that may inform quality improvement (QI) work (e.g. among patient transferred in from particular facilities).

The Hospital-Specific Reports also provide hospitals with more detailed benchmarks with which to gauge their performance relative to peer hospitals and interpret their results, including comorbidity frequencies for their patients relative to other hospitals in their state and the country.

Additionally, the code used to process the claims data and calculate measure results is written in SAS (Cary, NC) and is provided each year to hospitals upon request.

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.

In April of each year, hospitals have access to the following list of updated resources related to the measure which is provided directly or posted publicly for hospitals to use:

1. Hospital-Specific Reports (HSR): available for hospitals to download from QualityNet in April of each calendar year; includes information on the index admissions included in the measure calculation for each facility, detailed measure results, and state and national results.
2. HSR User Guide: available with the HSR and posted on QualityNet; provides instructions for interpreting the results and descriptions of each data field in the HSR.
3. Mock HSR: posted on QualityNet; provides real national results and simulated state and hospital results for stakeholders who do not receive an HSR.
4. IQR Preview Reports and Preview Report Help Guide: available for hospitals to download from QualityNet in April of each calendar year; includes measure results that will be publicly reported on Hospital Compare.
5. Annual Updates and Specification Reports: posted in April of each calendar year on QualityNet with detailed measure specifications, descriptions of changes made to the measure specifications with rationale and impact analysis (when appropriate), updated risk variable frequencies and coefficients for the national cohort, and updated national results for the new measurement period.
6. Frequently asked Questions (FAQs): includes general and measure-specific questions and responses, as well as infographics that explain complex components of the measure’s methodology, and are posted in April of each calendar year on QualityNet.
7. The SAS code used to calculate the measure with documentation describing what data files are used and how the SAS code works. This code and documentation are updated each year and are released upon request beginning in July of each year.
8. Measure Fact Sheets: provides a brief overview of measures, measure updates, and are posted in April of each calendar year on QualityNet.

In July of each year, the publicly-reported measure results are posted on Hospital Compare, a tool to find hospitals and compare their quality of care that CMS created in collaboration with organizations representing consumers, hospitals, doctors, employers, accrediting organizations, and other federal agencies. Measure results are updated in July of each calendar year.

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.

Questions and Answers (Q&A)

The measured entities (acute care hospitals) and other stakeholders or interested parties submit questions or comments about the measure through an email inbox (CMSmortalitymeasures@yale.edu). Experts on measure specifications, calculation, or implementation, prepare responses to those inquiries and reply directly to the sender. We consider issues raised through the Q&A process about measure specifications or measure calculation in measure reevaluation.

Literature Reviews

In addition, we routinely scan the literature for scholarly articles describing research related to this measure. We summarize new information obtained through these reviews every 3 years as a part of comprehensive reevaluation as mandated by the Measure Management System (MMS) Blueprint.

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

Summary of Questions or Comments from Hospitals submitted through the Q & A process:

For the CABG mortality measure inquiries received from hospitals since the submission of the last annual form in December 2016 have included the following:

1. Requests for detailed measure specifications including ICD-9 and ICD-10 codes used to define the measure cohort or in the risk-adjustment model;
2. Requests for the SAS code used to calculate measure results;
3. Requests for clarification of how inclusion and exclusion criteria are applied, such as if hospice patients are included in the measure cohort;
4. Requests for hospital-specific measure information, such as data included in the HSRs; and
5. Requests for interpretation and clarification of results.

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

Summary of Question and Comments from Other Stakeholders:

For the CABG mortality measure, feedback received from other stakeholders since the submission of the last annual form in December 2016 has included the following:

1. Requests for detailed measure specifications including ICD-9 and ICD-10 codes used to define the measure cohort or in the risk-adjustment model;
2. Requests for the SAS code used to calculate measure results; and
3. Requests for clarification of how inclusion and exclusion criteria are applied, such as if hospice patients are included in the measure cohort;

Summary of Relevant Publications from the Literature Review:

Since December 2015, we have reviewed 8 articles related to mortality following isolated CABG surgery. One article examined CABG mortality disparities using a similarly measure, but no articles employed the measure score results.

References:

1. RH Mehta, DM Shahian, S Sheng, SM O'Brien, FH Edwards, JP Jacobs and ED Peterson. Association of Hospital and Physician Characteristics and Care Processes with Racial Disparities in Procedural Outcomes Among Contemporary Patients Undergoing Coronary Artery Bypass Grafting Surgery. Circulation. 2016;133:124-130, http://dx.doi.org/10.1161/CIRCULATIONAHA.115.015957

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.

Each year issues raised through the Q&A or in the literature related to this measure are considered by measure and clinical experts. Any issues that warrant additional analytic work due to potential changes in the measure specifications are addressed as a part of annual measure reevaluation. If small changes are indicated after additional analytic work is complete, those changes are usually incorporated into the measure in the next measurement period. If the changes are substantial CMS may propose the changes through rulemaking and adopt the changes only after CMS received public comment on the changes and finalizes those changes in the IPPS or other rule. There were no questions or issues raised by stakeholders requiring additional analysis or changes to the measures since the last annual form submission. There have been no changes made to the CABG measure since the time of its initial 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 median hospital 30-day, all-cause, RSMR for the CAGB mortality measure for the 3-year period between July 1, 2013 and June 30, 2016 was 3.1%. The median RSMR decreased by 0.1 absolute percentage points from July 2013-June 2014 (median RSRR: 3.1%) to July 2015-June 2016 (median: RSRR: 3.0%).

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.

N/A

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

N/A

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)

0114 : Risk-Adjusted Postoperative Renal Failure

0115 : Risk-Adjusted Surgical Re-exploration

0119 : Risk-Adjusted Operative Mortality for CABG

0122 : Risk-Adjusted Operative Mortality for Mitral Valve (MV) Replacement + CABG Surgery

0123 : Risk-Adjusted Operative Mortality for Aortic Valve Replacement (AVR) + CABG Surgery

0129 : Risk-Adjusted Postoperative Prolonged Intubation (Ventilation)

0130 : Risk-Adjusted Deep Sternal Wound Infection

0131 : Risk-Adjusted Stroke/Cerebrovascular Accident

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

0468 : Hospital 30-day, all-cause, risk-standardized mortality rate (RSMR) following pneumonia hospitalization

0535 : 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

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

1502 : Risk-Adjusted Operative Mortality for Mitral Valve (MV) Repair + CABG Surgery

1893 : Hospital 30-Day, all-cause, risk-standardized mortality rate (RSMR) following chronic obstructive pulmonary disease (COPD) hospitalization

2515 : Hospital 30-day, all-cause, unplanned, risk-standardized readmission rate (RSRR) following coronary artery bypass graft (CABG) surgery

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.

We did not include in our list of related measures any non-outcome (e.g., process) measures with the same target population as our measure. Our measure cohort was heavily vetted by clinical experts, a technical expert panel, and a public comment period. In addition, the related claims-based CABG readmission measure, which utilizes the same definition of isolated CABG as the mortality measure, was validated using STS clinical registry data. Because this is an outcome measure, clinical coherence of the cohort takes precedence over alignment with related non-outcome measures. Furthermore, non-outcome measures are limited due to broader patient exclusions. This is because they typically only include a specific subset of patients who are eligible for that measure (for example, patients who receive a specific medication or undergo a specific procedure).

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 NQF-endorsed STS measure that has the same target population and similar measure focus as the proposed CABG mortality measure is the Risk-adjusted operative mortality for CABG (NQF #0119). The measure steward for the registry-based mortality measure for CABG is STS. In developing the measure, we sought to harmonize with the STS measure to the greatest extent feasible given competing measure design objectives and differences in the data source. The potential sources of discrepancy are target patient population, age, isolated CABG, period of observation, and included hospitals. The STS measure also assesses both deaths occurring during CABG hospitalization (in-hospital death, even if after 30 days) and deaths occurring within 30 days of procedure date. As indicated above, the proposed measure uses a standard follow-up period of 30 days of procedure date in order to measure each patient consistently. The proposed claims-based measure has been tested and is appropriate for use in all-payer data for patients 18 years and over. Finally, the STS cardiac surgery registry currently enrolls most, but not all, patients receiving CABG surgeries in the U.S. The proposed

CABG mortality measure will capture all qualifying Medicare FFS patients undergoing CABG regardless of whether their hospital or surgeon participates in the STS registry.

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

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 (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.

Technical Expert Panel Members:

Joseph V. Agostini, MD, Aetna

Tanya Alteras, MPP, National Partnership for Women and Families

Mary Barton, MD, MPP, National Committee for Quality Assurance (NCQA)

Carol Beehler, RN, NEA-BC, Pricewaterhouse Coopers

Todd Michael Dewey, MD, Southwest Cardiothoracic Surgeons

Lee Fleisher, MD (Served from March 30, 2012 to May 25, 2012), American Society of Anesthesiologists, University of Pennsylvania School of Medicine

Paul Kurlansky, MD, Florida Heart Research Institute, Inc

Frederic Masoudi, MD, MSPN, University of Colorado-Denver, Senior Medical Office of National CV Data Registries

Christine McCarty, MD, Cardiovascular Surgical Institute

Joseph Parker, PhD, State of California: Office of Statewide Health Planning and Development,

Kenneth Sands, MD, MPH, Beth Israel Deaconess Medical Center

Ed Savage, MD, Cleveland Clinical Florida

Stephen Schmaltz, PhD, The Joint Commission

Richard Shemin, MD, UCLA Medical Center

Alan Speir, MD, Inova Fairfax Hospital

Working Group Panel Members:

Arnar Geirsson, MD, Yale School of Medicine

David Shahian, MD, STS Workforce on National Databases, Harvard Medical School, Massachusetts General Hospital

Measure Developer/Steward Updates and Ongoing Maintenance

Ad.2 Year the measure was first released: 2014

Ad.3 Month and Year of most recent revision: 12, 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, 2019

Ad.6 Copyright statement: N/A

Ad.7 Disclaimers: N/A

Ad.8 Additional Information/Comments: N/A