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: Ctrl + click link to go to the link; ALT + LEFT ARROW to return

Brief Measure Information

NQF #: 3504

Measure Title: Claims-Only Hospital-Wide (All-Condition, All-Procedure) Risk-Standardized Mortality Measure

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

Brief Description of Measure: The measure estimates a hospital-level 30-day hospital-wide risk-standardized mortality rate (RSMR), defined as death from any cause within 30 days after the index admission date, for Medicare fee-for-service (FFS) patients who are between the ages of 65 and 94.

Please note that in parallel with the claims-only HWM measure, we are submitting a hybrid HWM measure. Note that ultimately the claims and hybrid measures will be harmonized and use the same exact cohort specifications. The intent is that prior to implementation, the two measures will be exactly the same, with the exception of the additional risk adjustment added by the CCDE in the hybrid measure. This is analogous to the currently endorsed and implemented hybrid hospital-wide readmissions measure (NQF 1789 and NQF 2879e).

Because of the homology between the claims and hybrid HWM measures, there is no reason to suspect that the results of analyses done for the claims-only measure would differ in any significant way from results of analyses for a nationally representative hybrid measure.

Below we highlight the differences between the two measures, including specifications, data used, and testing which reflect limitations of data availability, as well as actual intended differences in the measure (risk adjustment).

Differences in the measure, data, and testing that reflect limitations in data availability

1. Dataset used for development, some testing (see below for differences), and measure results:

a. The claims-only measure uses nation-wide Medicare FFS claims and the enrollment database.

b. The hybrid measure uses an electronic health record (EHR) database from 21 hospitals in the Kaiser Permanente network which includes inpatient claims data information.

2. Age of patients in cohort:

a. The claims-only measure includes Medicare FFS patients, age 65-94.

b. The hybrid measure includes all patients age 50-94 (see later discussion for justification)

3. External empiric validity testing

a. Not possible for the hybrid measure, due to limited data availability. We provide results from the claims-only measure within the hybrid testing form.

4. Socioeconomic risk factor analyses

a. Not possible for the hybrid measure, due to limited data availability. We provide results from the claims-only measure within the hybrid testing form.

5. Exclusion analyses

a. To be representative of what we expect the impact would be of the measures' exclusions in a nation-wide sample, we provide the results from the claims-only measure.

6. Meaningful differences

a. To be representative of what we expect the range of performance would be in a nation-wide sample, we provide the distribution results from the claims-only measure.

Difference between the two measures when fully harmonized, prior to implementation:

1. Risk adjustment:

a. The claims-only measure uses administrative claims data only for risk adjustment

b. The hybrid measure adds 10 clinical risk variables, derived from a set of core clinical data elements (CCDE) extracted from the EHR.

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 range of medical conditions and surgical procedures. 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 their patient case mix, and therefore promote hospital quality improvement and better inform consumers about care quality. The more granular division-level results can support the targeting of service-line quality improvement.

Mortality is a significant outcome that is meaningful to patients and providers. For the majority of Medicare beneficiaries admitted to acute care hospitals in the US, the goal is to avoid short-term mortality. According to recent internal analyses, from July 2016 to June 2017, there were about 10 million inpatient admissions among Medicare FFS beneficiaries between the ages of 65 and 94, at 4,700 US hospitals. The observed mean 30-day mortality rate was 8.17%. The range of mortality scores on the HWM measure from 4692 acute-care hospitals was 3.95%-8.70% across more than 4.3 million admissions.

Numerator Statement: The outcome for this measure is 30-day, all-cause mortality. Mortality is defined as death from any cause, either during or after admission, within 30 days of the index admission date.

Denominator Statement: The cohort includes inpatient admissions for a wide variety of conditions for Medicare FFS patients aged between 65 and 94 years old who were admitted to short-term acute care hospitals. If a patient has more than one admission during the measurement year, one admission is randomly selected for inclusion in the measure. Additional details are provided in S.7 Denominator Details.

Denominator Exclusions: The measure excludes index admissions for patients:

1. With inconsistent or unknown vital status (from claims data) or other unreliable claims data;

2. Discharged against medical advice (AMA);

3. With an admission for spinal cord injury (CCS 227), skull and face fractures (CCS 228), Intracranial Injury (CCS 233), Crushing injury or internal injury (CCS 234), Open wounds of head/neck/trunk (CCS 235), and burns (CCS 240); and

4. With a principal discharge diagnosis within a CCS with fewer than 100 admissions within the measurement year.

Measure Type: Outcome

Data Source: Claims, Enrollment Data, Other

Level of Analysis: Facility

Preliminary Analysis: New Measure

Criteria 1: Importance to Measure and Report

1a. Evidence

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

- The developer provided a logic model that outlines the relationship between various healthcare processes and interventions, improved health status and a decreased risk of mortality.
- The developer provided several evidence-based strategies to reduce hospital mortality:
    - Adoption of strategies shown to reduce ventilator-associated pneumonia
    - Delivery of reliable, evidence-based care for acute myocardial infarction
    - Prevention of adverse drug events though medication reconciliation
    - Prevention of central line infections through evidence-based guideline-concordant care
    - Prevention of surgical site infections through evidence-based guideline-concordant care
    - Use multidisciplinary rounds to improve communication

  - Employ Rapid Response Teams to attend to patients at the first sign of clinical decline
  - Identify high-risk patients on admission and increase nursing care and physician contact accordingly
  - Standardize patient handoffs to avoid miscommunication or gaps in care
  - Establish partnerships with community providers to promote evidenced-based practices to reduce hospitalizations before patients become critically ill

Questions for the Committee:

- *Is there at least one thing that the provider can do to achieve a change in the measure results?*
- *Does this broadly-focused mortality measure provide meaningful information beyond the condition-specific mortality measures?*

Guidance from the Evidence Algorithm

Outcome measure (Box 1) → Relationship between outcome and at least one healthcare action demonstrated by data (Box 2) → Yes → 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

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

- Data from the claims-only version of the measure were provided, as the hybrid measure dataset was limited and not representative of the nation as a whole.
  - The hybrid HWM uses the same concept, cohort and outcome and claims-based risk adjustment variables as the claims-only measure. (It also incorporates clinical data elements from EHR data in the risk-adjustment model).
- From July 2016 to June 2017, there were about 10 million inpatient admissions among Medicare FFS beneficiaries between the ages of 65 and 94, at 4,700 US hospitals. The observed mean 30-day mortality rate was 8.17%.
- In the study cohort, the mean hospital-level risk standardized mortality rate (RSMR) was 6.85%. The range of mortality scores on the HWM measure from 4692 acute-care hospitals was 3.95%-8.70% across more than 4.3 million admissions.
- RSMR Distribution
  - Min, 3.95%
  - 1st, 5.57%
  - 5th, 6.07%
  - 10th, 6.32%
  - 25th, 6.66%
  - 50th, 6.93%

- ○ 75th, 7.09%
- ○ 90th, 7.26%
- ○ 95th, 7.40%
- ○ 99th, 7.75%
- ○ Max, 8.70%

- The developer also provides evidence from the literature that more than 400,000 patients die each year from preventable harm in hospitals and high and variable mortality rates across hospitals indicate opportunities for improvement.

Disparities

- The developer provides: a) The distribution of patients that were dual-eligible and those with low AHRQ SES (using claims-based measure and 2016-2017 dataset) across hospitals is provided and b) measure score percentiles (bottom vs. top quartile for proportion of patients with low SES and dual-eligibility).

Questions for the Committee:

- Is there a gap in care that warrants a national performance measure?
- Do disparites exist that warrant consideration of stratification or adjustment for social factors?

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

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

1a. Evidence

Comments:

**appropriate evidence

**Solid evidence

1b. Performance Gap

Comments:

**gap noted , disparities also noted , risk adj might be warranted

**Yes there is a gap.

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability: Specifications **and** Testing

2b. Validity: Testing; Exclusions; Risk-Adjustment; Meaningful Differences; Comparability Missing Data

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

Reliability

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

2a2. Reliability testing demonstrates if the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise enough to distinguish differences in performance across providers. For maintenance measures – less emphasis if no new testing data provided.

Validity

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

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

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

Evaluators: NQF Scientific Methods Panel Subgroup

Methods Panel Review (Combined)

Scientific Methods Panel Votes: Measure passes

- Reliability: H-3, M-2, L-0, I-0
- Validity: H-3, M-2, L-0, I-0

This measure was reviewed by the Scientific Methods Panel and discussed on their call. A summary of the measure is provided below:

Reliability

- ○ Reliability testing was performed for measure score
- ○ Measure Score
    - ○ To assess reliability of the measure score, the developer conducted a “split sample” analysis, calculating the intra-class correlation coefficient between scores (i.e., RSMRs) generated from two randomly selected subsets from each hospital.
    - ○ The agreement between the two independent assessments of the RSMR for each hospital was 0.8187, and the adjusted ICC (which estimates the ICC if the developer had been able to use one full year of data in each split sample) [3,4], is 0.8377 (Table 1). Both demonstrate high reliability, according to conventional standards.

Validity

- ○ Validity was tested at the measure score level. Both empirical validity testing and a face validity assessment were provided. NQF clarified that for this new measure, face validity alone is sufficient.

- Meaningful differences between hospitals are illustrated in Table 11 and Figure 9, risk adjusted hospital performance is proffered in percentile tabulation and histogram. Range from worst (<1%tile) to best (>99%tile), is 3.95% to 8.70% so scores fall in a relatively tight range.
- Empirical Measure Score
  - To test the validity of the measure score, the developer examined whether better performance on the measure was related to better performance for other relevant structural and outcome measures.
  - The measure score was correlated with the following three measures:
    - Nurse-to-bed ratio
    - Hospital star rating mortality group score
    - Overall hospital star rating
  - The developer reports that, for each external measure of quality, the comparison showed a trend toward better performance on the HWM measure with better performance on the comparator measure; detailed results are provided in a series of charts in the testing form.
- Face Validity
  - A total of 6 TEP members completed the face validity survey.
  - Of the 6 respondents, 5 respondents (83%) indicated that they somewhat, moderately, or strongly agreed, and 1 moderately disagreed that the claims-based measure can be used to distinguish between better and worse quality facilities.
- The measure uses a statistical risk model with 21 factors.
  - Social risk factors, dual-eligible status and AHRQ SES, were tested but not included in the model.

Standing Committee Action Item(s): The Standing Committee can discuss reliability and/or validity or accept the Scientific Methods Panel ratings.

Questions for the Committee regarding reliability:

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

Questions for the Committee regarding validity:

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

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

Methods Panel Evaluation (Combined): Scientific Acceptability

Scientific Acceptability: Preliminary Analysis Form

Measure Number: 3504

Measure Title: Insert measure title here

Type of measure:

☐ **Process** ☐ **Process: Appropriate Use** ☐ **Structure** ☐ **Efficiency** ☐ **Cost/Resource Use**

☒ **Outcome** ☐ **Outcome: PRO-PM** ☐ **Outcome: Intermediate Clinical Outcome** ☐ **Composite**

Data Source:

☒ **Claims** ☒ **Electronic Health Data** ☐ **Electronic Health Records** ☐ **Management Data**
☐ **Assessment Data** ☐ **Paper Medical Records** ☐ **Instrument-Based Data** ☐ **Registry Data**
☒ **Enrollment Data** ☒ **Other** (Medicare enrollment data; hospice data file)

Level of Analysis:

☐ **Clinician: Group/Practice** ☐ **Clinician: Individual** ☒ **Facility** ☐ **Health Plan**
☐ **Population: Community, County or City** ☐ **Population: Regional and State**
☐ **Integrated Delivery System** ☐ **Other**

Measure is:

☒ **New** ☐ **Previously endorsed (**NOTE: Empirical validity testing is expected at time of maintenance review; if not possible, justification is required.)

Methods Panel member 1: Note 1 - this section is blank. Note 2—this measure (Hybrid; Kaiser Permanente only) has a "sister" measure (claims only; general Medicare FFS-based). The developer lists several differences (e.g., risk adjustment methodology and RFs; patient age group) between these measures in section De.3. The intent is to harmonize the two measures prior to implementation.

RELIABILITY: SPECIFICATIONS

1. **Are submitted specifications precise, unambiguous, and complete so that they can be consistently implemented?** ☒ **Yes** ☐ **No**

   Submission document: "MIF_xxxx" document, items S.1-S.22

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

2. **Briefly summarize any concerns about the measure specifications.**

   Methods Panel member 2: The MIF form was not available, so I can't be sure of this, but there do not seem to be any concerns.

   Methods Panel member 1: The measure is restricted to hospitals that have at least 100 "admissions in that division". This will restrict reportability of the measure and may exclude smaller and rural hospitals. If this measure is used in value-based purchasing, some recognition of benefits achieved by hospitals for which the measure can be computed should be awarded to those hospitals for which the measure cannot be computed.

Developer Response, 5/22/2019: One of the main goals of the HWM measure is to provide hospital quality information for smaller, low volume hospitals. In order to receive an overall risk-standardized mortality rate, the hospital must have at least 25 cases overall. 4,455 out of the total 4,692 (95%) have at least 25 cases and would have received a HWM score using July 2016 – June 2017 data. We exclude admissions within low volume CCSs, defined as less than or equal to 100 patients within any division, for more stable, precise division-level risk models. This exclusion is on the CCS level (using total count from all hospitals), not on the individual hospital-level.

RELIABILITY: TESTING

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

3. **Reliability testing level** ☒ **Measure score** ☒ **Data element** ☐ **Neither**

4. **Reliability testing was conducted with the data source and level of analysis indicated for this measure** ☒ **Yes** ☐ **No**

5. If score-level and/or data element reliability testing was NOT conducted or if the methods used were NOT appropriate, was **empirical VALIDITY testing** of patient-level data conducted?

   ☐ **Yes** ☐ **No**

   Methods Panel member 1: N/A - reliability testing was conducted at both the measure score and data element levels

6. **Assess the method(s) used for reliability testing**

   Submission document: Testing attachment, section 2a2.2

   Methods Panel member 2: A split sample approach was used in which two independent samples were drawn and each used to create a RSMR for each hospital, which were then compared with an ICC.

   Methods Panel member 3**:** Tested reliability of the measure with a split sample approach.

   Methods Panel member 1: For the measure score, a split-half analysis (intra class correlation—ICC)was computed. There was no information about the data element level reliability—although that box was checked on Section 2a2.1. The data analyzed were collected over 15 months, while the reporting timeline for the measure appears to be for a 12-month period.

   Developer Response, 5/22/2019: The hybrid HWM measure used 15 months. The claims-only HWM measure used 21 months of data for split-sample testing. To reliably calculate ICC for the split-sample, 24 months of data was required in total (one year for each split sample), however we were only able to obtain 21 months of data. Therefore, the ICC was then adjusted so that the results would project what the result would have looked like with 24 months of data. The measure is intended to be eventually publicly reported using a 12-month period, as more data would be available in the national sample.

7. **Assess the results of reliability testing**

   Submission document: Testing attachment, section 2a2.3

   Methods Panel member 2: The overall split-sample reliability score (0.8377) is indicates a high level of reliability.

   Methods Panel member 4: Reliability was assessed using split-sample reliability testing. ICC – 0.82, which is consistent with high reliability.

   Methods Panel member 3**:** Calculated a ICC of 0.8187 – indicating high reliability.

   Methods Panel member 1: Measure score split-half reliability = 0.8187 was substanial. The "adjusted ICC" value was higher, but methodology for how value was adjusted was not specified. Number of hospitals was sufficient (n=4450).

   There were no results posted for the data elements.

Developer Response, 5/22/2019: To calculate ICC, admissions were randomly and evenly split into the two split samples within each individual hospital (21 months of combined data). For each sample, we fit a hierarchical generalized linear model for each service-line division and then aggregate the results into an overall RSMR. That is, each hospital will have a RSMR in each split sample. The ICC evaluates the agreement between the RSMR calculated in the two randomly split samples. The adjusted ICC adjusts the time window from 21 months to 24 moths based on the 'Spearman Brown prophecy formula', which suggests a way to estimate the reliability of the measure if the whole cohort were used, based on an estimate from half the cohort. Specifically, the formula says that if the number of items in a test increases by a factor of N, then the new reliability P' can be estimated from the original reliability P using: $\rho' = N*\rho/(1+ (N-1)*\rho)$ where N=24/21=8/7

We did not provide or claim to provide data element reliability testing since we provided measure score reliability testing.

8. 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.

   Submission document: Testing attachment, section 2a2.2

   ☒ **Yes**

   ☐ **No**

   ☐ **Not applicable** (score-level testing was not performed)

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

   Submission document: Testing attachment, section 2a2.2

   ☐ **Yes**

   ☐ **No**

   ☒ **Not applicable** (data element testing was not performed)

10. **OVERALL RATING OF RELIABILITY** (taking into account precision of specifications 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** (NOTE: Should rate LOW if you believe specifications are NOT precise, unambiguous, and complete or if testing methods/results are not adequate)

☐ **Insufficient** (NOTE: Should rate INSUFFICIENT if you believe you do not have the information you need to make a rating decision)

11. **Briefly explain rationale for the rating of OVERALL RATING OF RELIABILITY and any concerns you may have with the approach to demonstrating reliability.**

Methods Panel member 2: An appropriate method was used, with strong empirical results.

Methods Panel member 5: ICC, moderate

Methods Panel member 1: I only award "High" if both measure score and data element reliability are demonstrated with strong results.

Methods Panel member 4: Reliability was assessed using split-sample reliability testing. ICC – 0.82, which is consistent with high reliability.

Methods Panel member 3**:** Conducted score-level reliability testing; used appropriate method; ICC of 0.8187 indicates high reliability.

VALIDITY: ASSESSMENT OF THREATS TO VALIDITY

12. **Please describe any concerns you have with measure exclusions.**

Submission document: Testing attachment, section 2b2.

Methods Panel member 2: None.

Methods Panel member 5: N/A

Methods Panel member 3**:** No concerns.

Methods Panel member 4: No concerns. 0.21% of cases were excluded in order to only include CCS codes with >100 observations in order for risk adj model to converge. Low-volume CCS categories were not combined into single category due to heterogeneity of this combined group.

Methods Panel member 1: See previous comment under measure description.

13. **Please describe any concerns you have regarding the ability to identify meaningful differences in performance.**

Submission document: Testing attachment, section 2b4.

Methods Panel member 3**:** No concerns

Methods Panel member 2: None.

Methods Panel member 5: None.

Methods Panel member 4: None. The RSMR was 6.3% for 10th percentile and 7.26% for 90th percentile, indicating a meaningful quality gap.

Methods Panel member 1: Between hospital meaningful differences were not specifically described; no results presented.

14. **Please describe any concerns you have regarding comparability of results if multiple data sources or methods are specified.**
Submission document: Testing attachment, section 2b5.
Methods Panel member 5: N/A
Methods Panel member 3**:** Not applicable.
Methods Panel member 1: Data from FFS hospitals seem appropriate.

15. **Please describe any concerns you have regarding missing data.**
Methods Panel member 2: None.

Submission document: Testing attachment, section 2b6.

Methods Panel member 5: None

Methods Panel member 3**:** Not applicable.
Methods Panel member 1: No missing data.

16. **Risk Adjustment**

16a. **Risk-adjustment method** ☐ **None** ☒ **Statistical model** ☐ **Stratification**

16b. **If not risk-adjusted, is this supported by either a conceptual rationale or empirical analyses?**

☐ Yes ☐ No ☒ Not applicable

16c. **Social risk adjustment:**

16c.1 Are social risk factors included in risk model? ☒ Yes ☐ No ☐ Not applicable

16c.2 Conceptual rationale for social risk factors included? ☒ Yes ☐ No

16c.3 Is there a conceptual relationship between potential social risk factor variables and the measure focus? ☒ Yes ☐ No

Methods Panel member 4: Correlation coefficient for hospital RSMR with and without inclusion of patient-level SES (duals) was 0.999 indicating little effect of adjusting for patient SES on overall hospital performance.

16d.**Risk adjustment summary:**

16d.1 All of the risk-adjustment variables present at the start of care? ☒ Yes ☐ No
16d.2 If factors not present at the start of care, do you agree with the rationale provided for inclusion? ☐ Yes ☒ No (N/A) - 21 risk factors in model
16d.3 Is the risk adjustment approach appropriately developed and assessed? ☒ Yes ☐ No
16d.4 Do analyses indicate acceptable results (e.g., acceptable discrimination and calibration) ☐ Yes ☐ No
16d.5.Appropriate risk-adjustment strategy included in the measure? ☐ Yes ☐ No

16e**. Assess the risk-adjustment approach**

Methods Panel member 2: Overall, very thoughtful and through.

Methods Panel member 3**:** Model includes 21 risk factors, which are all present at the start of care; c-statistics were good.

Methods Panel member 4: A separate risk adjustment model was specified for each of the 15 mutually exclusive surgical and medical diagnostic groups using hierarchical logistic regression modelling. Each model adjusts for 21 risk factors based on the CMS HCC and principal discharge diagnoses based on AHRQ CCS. Provider performance is quantified using the PE ratio. Hopsitals were specified as a random effect. The PE ratio was estimated for each diagnostic group separately, and these PE ratios were then pooled for each hospital using an inverse variance-weighted geometric mean to create a hospital-wide composite PE ratio.

Model discrimination and calibration were acceptable in the validation data set for each of the component risk adj models, ranging between 0.75 to 0.91. Overfit stat were consistent with excellent calibration.

Methods Panel member 1: Prediction model for risk adjustment seem appropriate, and with acceptable results.

VALIDITY: TESTING

17. **Validity testing level:** ☒ **Measure score** ☐ **Data element** ☐ **Both**
18. **Method of establishing validity of the measure score:**

☒ **Face validity**

☒ **Empirical validity testing of the measure score**

☐ **N/A (score-level testing not conducted)**

19. **Assess the method(s) for establishing validity**

Submission document: Testing attachment, section 2b2.2

Methods Panel member 4: See below

Methods Panel member 1: The Developer conducted a number of very good predictive validity tests, as well as a face validity test using a TEP.

Methods Panel member 2: Face validity was assessed by surveying members of a Technical Experts Panel (TEP).

Empirical validity testing involved exploring the relationship of the HWM claims-only measure scores with each of the three external measures of hospital quality identified by the TEP: nurse-to-bed ratio, mortality group score of Star Rating, and Overall Star Rating. In addition, the developers identified "better" and "worse" outliers based on the 95% confidence interval.

Methods Panel member 3**: For the measure level score, used both face validity and empirical validity** testing. Face validty included a TEP that was about the usefulness of the measure for distinguishing quality. Empirical validity testing looked at performance on this measure vs. 3 other measures.

20. **Assess the results(s) for establishing validity**

Submission document: Testing attachment, section 2b2.3

Methods Panel member 4: See below

Methods Panel member 1: Results from predictive validity and face validit tests were very positive and produced consistent results.

Methods Panel member 2: The TEP results indicated relatively high agreement (83%) regarding the face-validity of the claims-only HWM measure.

For each external measure of quality, the comparison showed a trend toward better performance on the HWM measure with better performance on the comparator measure. For instance, In the outlier analysis that compares RSMR to nurse-to-bed ratio, there are 19 "better than national average" outliers in the third quartile and 5 "worse" outliers.

Methods Panel member 3**:** Face validity – 5/6 of TEP agreed that the measure does reflect quality

Empirical validity testing – hospital performance on this measure moves as expected with the other measures

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

    Submission document: Testing attachment, section 2b1.

    ☒ **Yes**

    ☐ **No**

    ☐ **Not applicable** (score-level testing was not performed)

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.*

    Submission document*: Testing attachment, section 2b1.*

    ☒ **Yes**

    ☐ **No**

    ☐ **Not applicable** (data element testing was not performed)

23. **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** (NOTE: Should rate LOW if you believe that there are threats to validity and/or relevant threats to validity were not assessed OR if testing methods/results are not adequate)

    ☐ **Insufficient** (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.)

24. **Briefly explain rationale for rating of OVERALL RATING OF VALIDITY and any concerns you may have with the developers' approach to demonstrating validity.**

Methods Panel member 4: Face validity was assessed by a TEP and, based on 6 of 8 TEP members, showed that 83% agreed with "The risk-standardized hospital mortality rates obtained from the claims-only HWM measure, as specified, can be used to distinguish between better and worse quality facilities."

Empiric validity was tested by comparing performance on HWM measure to 3 external measures of hospital quality: nurse-to-bed ratio, mortality group score of Star Rating, and Overall Star Rating. Results show better performance on HWM measure is associated with better performance on all 3 of the comparator measures. As acknowledged by MD, this type of analysis is always limited by the fact that there is no gold standard for quality.

Predictive validity was demonstrated by the acceptable performance of the risk adjustment model. Model discrimination and calibration were acceptable in the validation data set for each of the component risk adj models, ranging between 0.75 to 0.91. Overfit stat were consistent with excellent calibration.

Methods Panel member 1: I only award "High" if both measure score and data element validity are demonstrated with strong results.

Methods Panel member 2: Appropriate methods were used to assess both face validity and to conduct empirical testing, and the results were strong.

ADDITIONAL RECOMMENDATIONS

25. **If you have listed any concerns in this form, do you believe these concerns warrant further discussion by the multi-stakeholder Standing Committee? If so, please list those concerns below.**

Methods Panel member 1: Good job by developer.

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

2a1. Reliability – Specifications

Comments:

**No concerns

**No concerns

2a2. Reliability – Testing

Comments:

**no concerns

**No

2b1. Validity –Testing

Comments:

**No concerns

**No

2b4-7. Threats to Validity
2b4. Meaningful Differences

Comments:

**No concerns

**No concerns

2b2-3. Other Threats to Validity
2b2. Exclusions
2b3. Risk Adjustment

Comments:

**No concerns

**No concerns

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.

- All data elements are in defined fields of electronic sources and coded by someone other than person obtaining original information (e.g., DRG, ICD-9 codes on claims)

Questions for the Committee:

- Are the required data elements routinely generated and used during care delivery?
- Are the required data elements available in electronic form, e.g., EHR or other electronic sources?
- Is the data collection strategy ready to be put into operational use?

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

Committee Pre-evaluation Comments:
Criteria 3: Feasibility

3. Feasibility

Comments:

**No concerns

**In use today, no issues

Criterion 4: Usability and Use

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

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

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

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

Current uses of the measure

Publicly reported? ☐ **Yes** ☒ **No**

Current use in an accountability program? ☐ **Yes** ☒ **No** ☐ **UNCLEAR**

OR

Planned use in an accountability program? ☒ **Yes** ☐ **No**

Accountability program details

- Planned use in the Hospital Inpatient Quality Reporting Program.
- In the IPPS proposed rule, CMS signaled the eventual possibility of including this measure (and/or the related hybrid measure) within the Inpatient Quality Reporting (IQR) program.

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 measure is not yet implemented, but primary goal of the measure is to provide information necessary to implement focused quality improvement efforts.
- Developer notes plan to examine trends in improvements by comparing RSMRs over time.

Additional Feedback:

- N/A

***Questions for the Committee**:*

- Does the developer provide enough information regarding the potential use of this measure as well as how users can use results and provide feedback?
- Can performance results be used to further the goal of high-quality, efficient healthcare?
- How has the measure been vetted in real-world settings by those being measured or others?

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

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

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

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

Improvement results

- No additional information provided. (Performance variation provided in Opportunity for Improvement section, based on one year of data).

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

- Measure is not yet implemented.

Potential harms

- No potential harms identified.

Additional Feedback:

- N/A

Questions for the Committee:

- How can the performance results be used to further the goal of high-quality, efficient healthcare?
- Would there be potential harms from implementing this measure?
- Do the benefits of the measure outweigh any potential unintended consequences?

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

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

4a1. Use - Accountability and Transparency

Comments:

**No concerns

**No concerns

4b1. Usability – Improvement

Comments:

**No concerns

**No concerns

Criterion 5: Related and Competing Measures

Related or competing measures

- NQF 1789: Hospital-Wide All-Cause Risk-Standardized Readmission Measure
- NQF 1550: Hospital-level risk-standardized complication rate (RSCR) following elective primary total hip arthroplasty (THA) and/or total knee arthroplasty (TKA
- NQF 0468: Hospital 30-day, all-cause, risk-standardized mortality rate (RSMR) following pneumonia hospitalization
- NQF 1893: Hospital 30-Day, all-cause, risk-standardized mortality rate (RSMR) following chronic obstructive pulmonary disease (COPD) hospitalization
- NQF 2558: Hospital 30-Day, all-cause, risk-standardized mortality rate (RSMR) following coronary artery bypass graft (CABG) Surgery
- NQF 0230: Hospital 30-day, all-cause, risk-standardized mortality rate (RSMR) following acute myocardial infarction (AMI) hospitalization
- NQF 0229: Hospital 30-day, all-cause, risk-standardized mortality rate (RSMR) following heart failure (HF) hospitalization
- NQF 0347: Hospital 30-day, all-cause, risk-standardized mortality rate (RSMR) following acute ischemic stroke hospitalization. Death Rate in Low Mortality Diagnosis Related Groups (PSI-02)
- NQF 0530: AHRQ's Mortality for Select Conditions

Harmonization

- The developer states that the differences in the specifications are justified.

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

5. Related and Competing

Comments:

**related measures and not competing and developer will harmonize what they can before using this measure

**No

Public and Member Comments

Comments and Member Support/Non-Support Submitted as of: 6/5/2019

- **No NQF Members have submitted support/non-support choices as of this date.**

Developer Submission

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

Del19b1HOP5HWMClaimsEvidenceForm022819.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.

1a. Evidence (subcriterion 1a)

NATIONAL QUALITY FORUM—Evidence (subcriterion 1a)

Measure Number (*if previously endorsed*)**:** N/A

Measure Title: Claims-Only Hospital-Wide (All-Condition, All-Procedure) Risk-Standardized Mortality Measure

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

Date of Submission: N/A

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: Mortality

☐Patient-reported outcome (PRO): Click here to name the 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*): Click here to name the intermediate outcome

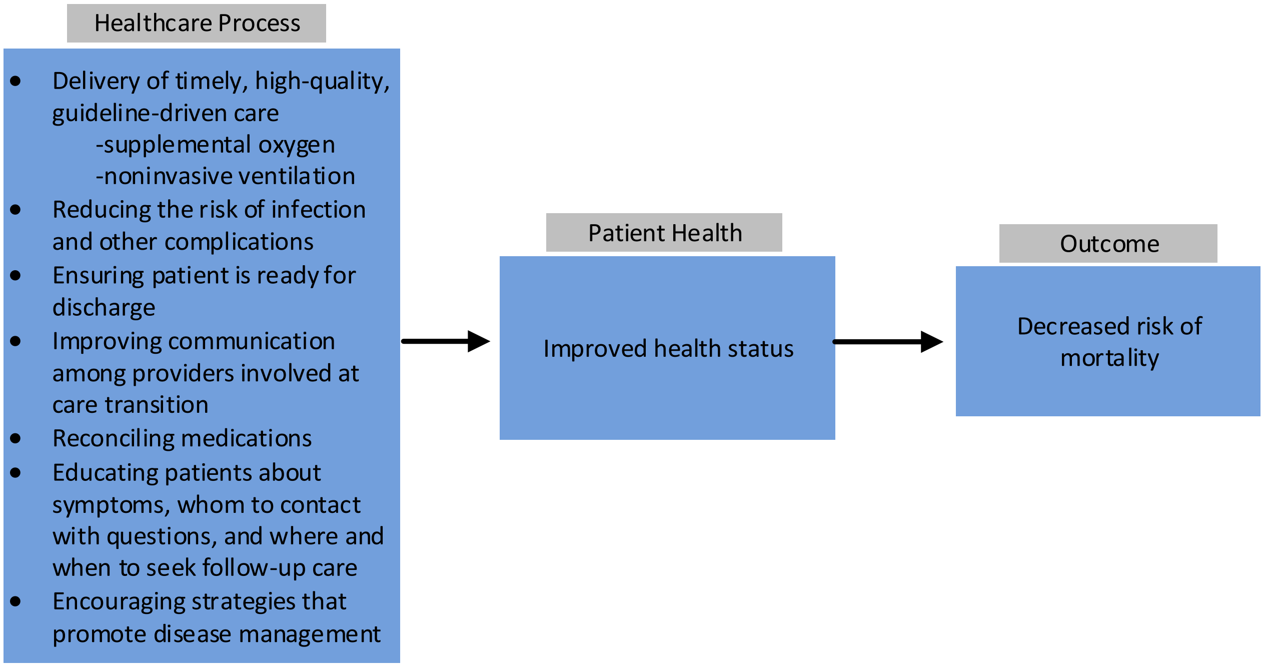
□ Process: Click here to name what is being measured

  □ Appropriate use measure: Click here to name what is being measured

□ Structure: Click here to name the structure

□ Composite: Click here to name what is being measured

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.



Healthcare Process

- Delivery of timely, high-quality, guideline-driven care
  - -supplemental oxygen
  - -noninvasive ventilation
- 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

Patient Health

Improved health status

Outcome

Decreased risk of mortality

The goal of this measure is to improve patient outcomes by providing patients, physicians, and hospitals with information about hospital-level, hospital-wide, risk-standardized mortality rates.

Mortality is an unwanted outcome for the overwhelming majority of patients admitted to US hospitals. Although mortality within 30 days of hospitalization is uncommon, when assessed among appropriate patients, it provides a concrete signal of care quality across conditions and procedures. It captures the result of care processes, such as peri-operative management protocols, and the impact of both optimal care and adverse events resulting from medical care.

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

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

****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.**

Death is a finite event, easy to measure accurately, and easily understood by patients and providers. For the majority of Medicare beneficiaries admitted to acute care hospitals in the US, the goal is to avoid short-term mortality. By measuring Hospital-Wide Mortality (HWM), CMS can ensure that efforts to reduce other outcomes, such as readmissions and resource utilization, are not resulting in unintended consequences. Specifically, this HWM measure will complement the existing CMS Hospital-Wide All-

Cause Risk-Standardized Readmission Measure (NQF #1789) to allow assessment of trends in hospital performance for both outcomes, similar to other complementary pairs of readmission and mortality measures for specific conditions and procedures. Further, the HWM measure will provide CMS with annually updated performance estimates for a larger proportion of the nation's hospitals, allowing significant performance outliers to be identified.

According to recent internal analyses, from July 2016 to June 2017, there were about 10 million inpatient admissions among Medicare Fee-for-Service (FFS) beneficiaries between the age of 65 and 94, at 4,700 US hospitals. The observed 30-day mortality rate was 8.17%. This is especially relevant as, while the current condition- and procedure-specific mortality measures address the most common and morbid healthcare conditions as identified by MedPAC[1] in the most recent three-year public reporting period, together they captured only 4.8 million Medicare FFS beneficiary admissions; a HWM measure is likely to capture about 6.5 million admissions across 4,700 hospitals. Using acute myocardial infarction as an example, which has seen the greatest declines in mortality, the median hospital risk-standardized mortality rate (RSMR) following admission for acute myocardial infarction has declined from 16.4% in 2006 to 13.1% in 2016 (July 2015-July 2016 data).[2,3] If development and reporting of this HWM measures produces even a tenth as much impact, this would translate into nearly 14,000 deaths averted in a one-year period. Furthermore, if all hospitals performed as well as hospitals in the 10th percentile for RSMR, about 100,000 deaths would be averted, compared to if all hospitals were performing at the median.

For some conditions and diagnoses, evidence supports that optimal medical care reduces mortality.[4,5] We know from ongoing improvements in condition- and procedure-specific mortality rates that interventions to improve these outcomes are feasible.[2] Multiple organizations, including the Institute for Healthcare Improvement (IHI), promote a range of evidence-based strategies to reduce hospital mortality.[6] These strategies include:

1. Adoption of strategies shown to reduce ventilator-associated pneumonia[7-9]
2. Delivery of reliable, evidence-based care for acute myocardial infarction[10,11]
3. Prevention of adverse drug events though medication reconciliation[12]
4. Prevention of central line infections through evidence-based guideline-concordant care[13]
5. Prevention of surgical site infections through evidence-based guideline-concordant care[14,15]

To reduce mortality, the IHI further encourages hospitals to use multidisciplinary rounds to improve communication, employ Rapid Response Teams to attend to patients at the first sign of clinical decline, identify high-risk patients on admission and increase nursing care and physician contact accordingly, standardize patient handoffs to avoid miscommunication or gaps in care, and establish partnerships with community providers to promote evidenced-based practices to reduce hospitalizations before patients become critically ill.[16] The IHI's 100,000 Lives Campaign, which was created to enlist hospitals in a coordinated effort to adopt the above interventions, led to an estimated more than 120,000 lives saved over the first 18 months of the campaign.[17]

Some of the evidence-based recommendations above apply to specific diagnoses. While condition- and procedure-specific initiatives to reduce mortality may broadly impact mortality rates across other conditions and procedures, there is likely more to be gained by a measure of hospital-wide mortality that can inform and encourage quality improvement efforts for patients not currently captured by existing CMS mortality measures. For example, a 2017 study of a standardized, inter-hospital transfer tool found that in-hospital mortality decreased for transferred patients following implementation of a one-page handover containing information critical for immediate patient care.[18]

In addition, there is evidence that a hospital's organizational culture is linked to key measures of hospital quality performance.[19] Since these cultural and leadership qualities affect the entire hospital, the claims-only HWM measure may provide important incentives for hospitals to not only examine their care processes and improve care for individual conditions, but may also provide incentives to encourage care transformation and improve overall organizational culture.

References:

1. MedPAC. March 2011 Report to the Congress: Medicare Payment Policy. 2011; http://www.medpac.gov/documents/reports/Mar11_EntireReport.pdf?sfvrsn=0 Accessed January 20, 2016

2. Medicare Hospital Quality Chartbook 2010 Performance Report on Outcomes Measures for Acute Myocardial Infarction, Heart Failure, and Pneumonia September 29, 2010. Prepared by Yale New Haven Health Services Corporation Center for Outcomes Research and Evaluation. Medicare Hospital Quality Chartbook 2010: https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/HospitalQualityInits/Downloads/HospitalChartBook.pdf Accessed February 27, 2019.

3. Trends in mortality rates following admission for acute myocardial infarction, chronic obstructive pulmonary disease, heart failure, pneumonia, and acute ischemic stroke. Prepared by Yale New Haven Health Services Corporation Center for Outcomes Research and Evaluation. Medicare Hospital Quality Chartbook 2017: https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/HospitalQualityInits/Downloads/2017-Chartbook.zip, Accessed February 27, 2019.

4. To Err Is Human: Building a Better Health System, Institute of Medicine (IOM), National Academy Press, Washington, DC (1999)

5. Classen DC, Resar R, Griffin F, et al. 'Global trigger tool'shows that adverse events in hospitals may be ten times greater than previously measured. Health affairs. 2011;30(4):581-589

6. Berwick DM, Calkins DR, McCannon CJ, Hackbarth AD. The 100,000 lives campaign: Setting a goal and a deadline for improving health care quality. *JAMA.* 2006;295(3):324-327.

7. Tablan O, Anderson L, Besser R, Bridges C, Hajjeh R. CDC; Healthcare Infection Control Practices Advisory Committee. Guidelines for preventing health-care-associated pneumonia, 2003: Recommendations of CDC and the Healthcare Infection Control Practices Advisory Committee. *MMWR Recommendation Reports.* 2004;53(RR-3):1-36.

8. American Thoracic Society, Infectious Diseases Society of America. Guidelines for the management of adults with hospital-acquired, ventilator-associated, and healthcare-associated pneumonia. *American Journal of Respiratory Critical Care Medicine.* 2005;171:388-416.

9. Resar R, Pronovost P, Haraden C, Simmonds T, Rainey T, Nolan T. Using a bundle approach to improve ventilator care processes and reduce ventilator-associated pneumonia. *Joint Commission Journal on Quality and Patient Safety.* 2005;31(5):243-248.

10. Antman EM, Anbe DT, Armstrong PW, et al. ACC/AHA guidelines for the management of patients with ST-elevation myocardial infarction; A report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines (committee to revise the 1999 guidelines for the management of patients with acute myocardial infarction). *Journal of the American College of Cardiology.* 2004;44(3):E1-e211.

11. Centers for Medicare & Medicaid Services. Hospital Quality Initiative Overview. 2005; http://www.allhealth.org/briefingmaterials/HospitalQualityInitiativeOverview-CMS-512.pdf. Accessed January 20, 2016.

12. Joint Commission. 2005 Joint Commission National Patient Safety Goals: Practical Strategies and Helpful Solutions for Meeting These Goals. 2005; http://teacherweb.com/NY/StBarnabas/Law-PublicPolicy/JCINT-2005.pdf. Accessed January 20, 2016.

13. O'Grady NP, Alexander M, Dellinger EP, et al. Guidelines for the prevention of intravascular catheter–related infections. *Clinical Infectious Diseases.* 2002;35(11):1281-1307.

14. Mangram AJ, Horan TC, Pearson ML, Silver LC, Jarvis WR, Committee HICPA. Guideline for prevention of surgical site infection, 1999. *American Journal of Infection Control.* 1999;27(2):97-134.

15. The Joint Commission. Surgical Care Improvement Project. 2005; http://www.jointcommission.org/surgical_care_improvement_project/. Accessed January 20, 2016.

16. Whittington J, Simmonds T, Jacobsen D. *Reducing hospital mortality rates.* Institute for Healthcare Improvement; 2005.

17. Poteliakhoff E. Update on IHI's 100k Lives Campaign. October 2006; http://hpm.org/us/c8/5.pdf. Accessed January 20, 2016.

18. Theobald CN, Choma NN, Ehrenfeld JM, Russ S, Kripalani S. Effect of a Handover Tool on Efficiency of Care and Mortality for Interhospital Transfers. *Journal of Hosp Medicine* 2017; 12(1):23-28.

19. Curry LA, Linnander EL, Brewster AL, Ting H, Krumholz HM, Bradley EH. Organizational culture change in U.S. hospitals: A mixed methods longitudinal intervention study. *Implementation Science.* 2015;10:29.

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.

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

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

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.

N/A

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

N/A

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

N/A

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

N/A

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 range of medical conditions and surgical procedures. 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 their patient case mix, and therefore promote hospital quality improvement and better inform consumers about care quality. The more granular division-level results can support the targeting of service-line quality improvement.

Mortality is a significant outcome that is meaningful to patients and providers. For the majority of Medicare beneficiaries admitted to acute care hospitals in the US, the goal is to avoid short-term mortality. According to recent internal analyses, from July 2016 to June 2017, there were about 10 million inpatient admissions among Medicare FFS beneficiaries between the ages of 65 and 94, at 4,700 US hospitals. The observed mean 30-day mortality rate was 8.17%. The range of mortality scores on the HWM measure from 4692 acute-care hospitals was 3.95%-8.70% across more than 4.3 million admissions.

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

We conducted analyses using data from July 2016 to June 2017 Medicare claims data (4,335,530 admissions from 4692 hospitals).

In the study cohort, the mean hospital-level risk standardized mortality rate (RSMR) was 6.85%, with a range of 3.95%-8.70%. As shown below, the median RSMR was 6.93% (25th and 75th percentiles were 6.66% and 7.09%, respectively).

RSMR Distribution:

Min, 3.95%

1st, 5.57%

5th, 6.07%

10th, 6.32%

25th, 6.66%

50th, 6.93%

75th, 7.09%

90th, 7.26%

95th, 7.40%

99th, 7.75%

Max, 8.70%

Below we provide RSMRs by decile of performance for the Overall RSMR

Decile: Min (RSMR)-Max (RSMR)

1: 3.95%-6.32%

2: 6.32%-6.58%

3: 6.58%-6.74%

4: 6.74%-6.85%

5: 6.85%-6.93%

6: 6.93%-6.99%

7: 6.99%-7.05%

8: 7.05%-7.13%

9: 7.13%-7.26%

10: 7.26%-8.70%

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.

Mortality is a significant outcome that is meaningful to patients and providers, and the vast majority of patients admitted to the hospital have survival as a primary goal. For the majority of Medicare beneficiaries admitted to acute care hospitals in the US, the goal is to avoid short-term mortality. According to recent internal analyses, from July 2016 to June 2017, there were about 10 million inpatient admissions among Medicare FFS beneficiaries between the ages of 65 and 94, at 4,700 US hospitals. The observed 30-day mortality rate was 8.17%.

Furthermore, estimates using data from 2002 to 2008 suggest that more than 400,000 patients die each year from preventable harm in hospitals.[1] While we do not expect mortality rates to be zero, studies have shown that mortality within 30 days of hospital admission is related to quality of care, and that high and variable mortality rates across hospitals indicate opportunities for improvement.[2,3]

In addition, hospital-wide mortality has been the focus of a number of previous quality reporting initiatives in the US and other countries. Prior efforts have met with some success and a number of challenges. Through our environmental scan and literature review, we identified multiple hospital-wide mortality measures reported at the state-level, and several at the health-system level. There is no hospital-wide mortality measure reported at the national-level in the United States.

While existing condition- and procedure-specific mortality measures provide specificity for targeted quality improvement work and may have contributed to national declines in hospital mortality rates for measured conditions (Suter et al., 2014), they do not, however, allow broader statements about a hospital's performance, nor do they meaningfully capture performance for small-volume hospitals. Further, existing mortality measures may not capture cross-cutting hospital-wide characteristics that also contribute to quality of care. These factors may be difficult to measure, such as a global culture of safety, good communication across teams, multidisciplinary care teams, coordination with community services and efforts, and effective care transitions.

References:

1. James JT. A new, evidence-based estimate of patient harms associated with hospital care. Journal of patient safety. 2013;9(3):122-128.

2. Peterson ED, Roe MT, Mulgund J, et al. Association between hospital process performance and outcomes among patients with acute coronary syndromes. JAMA. 2006;295(16):1912-1920.

3. Writing Group for the Checklist- I.C.U. Investigators, Brazilian Research in Intensive Care Network. Effect of a quality improvement intervention with daily round checklists, goal setting, and clinician prompting on mortality of critically ill patients: A randomized clinical trial. JAMA. 2016;315(14):1480-1490.

4. Suter LG, Li SX, Grady JN, et al. National patterns of risk-standardized mortality and readmission after hospitalization for acute myocardial infarction, heart failure, and pneumonia: update on publicly reported outcomes measures based on the 2013 release. Journal of general internal medicine. Oct 2014;29(10):1333-1340.

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.*

This analysis used Dataset 2, which included Medicare FFS claims from July 1, 2016 – June 30, 2017, with 4,692 hospitals and a total of 4,335,530 admissions (study cohort).

The distribution of patients with the dual eligible risk factor across measured hospitals was:

Median: 14.6%

Interquartile range: 9.2%-22.8%

Measure score (RSMR) percentile: bottom vs. top quartile for proportion of patients with dual eligible risk factor

Min: 4.28% vs. 4.50%

10th: 6.18% vs 6.21%

25th: 6.48% vs. 6.57%

Median: 6.73% vs. 6.8%

75th: 6.99% vs. 7.11%

90th: 7.31% vs. 7.44%

Maximum: 8.61% vs. 9.20%

The distribution of patients with the low SES AHRQ social risk factor across measured hospitals was:

Median: 17%

Interquartile range: 7.1%-34.2%

Measure score (RSMR) percentile: bottom vs. top quartile for proportion of patients with the low SES AHRQ social risk factor.

Min: 4.28% vs. 4.51%

10th: 6.22% vs 6.37%

25th: 6.55% vs. 6.61%

Median: 6.73% vs. 6.84%

75th: 7.03% vs. 7.19%

90th: 7.29% vs. 7.55%

Maximum: 8.37% vs. 9.20%

Please see section 2b3.4b of the testing form for an in-depth analysis of social risk factors.

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

2.3 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.

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

NATIONAL QUALITY FORUM—Measure Testing (subcriteria 2a2, 2b1-2b6)

Measure Number (*if previously endorsed*)**:** N/A
Measure Title: Claims-Only Hospital-Wide (All-Condition, All-Procedure) Risk-Standardized Mortality Measure
Date of Submission: 1/7/2019

Type of Measure:

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

Instructions

- **Measures must be tested for all the data sources and levels of analyses that are specified. *If there is more than one set of data specifications or more than one level of analysis, contact NQF staff* about how to present all the testing information in one form.**
- **For all measures, sections 1, 2a2, 2b1, 2b2, and 2b4 must be completed.**
- **For outcome and resource use measures, section 2b3 also must be completed.**
- **If specified for multiple data sources/sets of specificaitons (e.g., claims and EHRs), section 2b5 also must be completed.**
- **Respond to all questions as instructed with answers immediately following the question. All information on testing to demonstrate meeting the subcriteria for reliability (2a2) and validity (2b1-2b6) must be in this form. An appendix for *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.**
- **If you are unable to check a box, please highlight or shade the box for your response.**

- **Maximum of 25 pages (*incuding questions/instructions;* minimum font size 11 pt; do not change margins). *Contact NQF staff if more pages are needed.***
- **Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.**
- **For information on the most updated guidance on how to address social risk factors variables and testing in this form refer to the release notes for version 7.1 of the Measure Testing Attachment.**

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

2a2. Reliability testing [10] demonstrates the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise. For instrument-based measures (including PRO-PMs) and composite performance measures, reliability should be demonstrated for the computed performance score.

2b1. Validity testing [11] demonstrates that the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality. For instrument-based measures (including PRO-PMs) and composite performance measures, validity should be demonstrated for the computed performance score.

2b2. Exclusions are supported by the clinical evidence and are of sufficient frequency to warrant inclusion in the specifications of the measure; [12]

AND

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

2b3. For outcome measures and other measures when indicated (e.g., resource use):

- **an evidence-based risk-adjustment strategy (e.g., risk models, risk stratification) is specified; is based on patient factors (including clinical and social risk factors) that influence the measured outcome and are present at start of care; [14],[15] and has demonstrated adequate discrimination and calibration**

OR

- **rationale/data support no risk adjustment/ stratification.**

2b4. Data analysis of computed measure scores demonstrates that methods for scoring and analysis of the specified measure allow for identification of statistically significant and practically/clinically meaningful [16] differences in performance;

OR

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

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

2b6. Analyses identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and 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:	Measure Tested with Data From:

(***must be consistent with data sources entered in S.17***)	
☐ 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: Medicare enrollment data; hospice data file	☒ other: Medicare enrollment data, hospice data; US census data; Master Beneficiary Summary File (MBSF)

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 development and testing include Medicare Part A administrative inpatient claims and the Medicare Enrollment Database (EDB). We also used the EDB to identify patients who were enrolled in hospice. To assess socioeconomic factors, we used census as well as claims data (dual eligible status obtained through the Master Beneficiary Summary File (MBSF) Database; 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?

July 1, 2013 – June 30, 2015 for initial measure development and July 1, 2015-June 30, 2017 for testing with ICD-10 coded data; includes one year of inpatient claims to identify index hospitalizations, plus 12 months of history data for comorbidities. Please see section 1.7 for additional details.

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

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

☒ hospital/facility/agency	☒ hospital/facility/agency
☐ health plan	☐ health plan
☐ other: Click here to describe	☐ other: Click here to describe

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

For this measure, hospitals are the measured entities. All non-federal, acute care inpatient US hospitals (including territories) with Medicare fee-for-service (FFS) beneficiaries aged between 65 and 94 years old are included. The number of measured entities (hospitals) varies by testing type.

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.

Dataset	Description of Dataset	Use and Section in the Testing Attachment
Dataset #1: **Initial Development Dataset (ICD-9)**	Index dataset containing administrative inpatient hospitalization data, enrollment data, and post-discharge mortality status for Medicare FFS beneficiaries, 65 years and older on admission, hospitalized from July 1, 2014 – June 30, 2015. A history dataset that includes inpatient hospitalization data on each patient for the 12 months prior to the index admission was used to identify and select risk-adjustment variables. Non-Surgical Divisions Cancer: 38,635 Cardiac: 682,716 Gastrointestinal: 351,117 Infectious Disease: 555,864	• Section 2b3.3a Identification and selection of risk-adjustment variables

	Neurology: 267,384 Orthopedics: 131,747 Pulmonary: 548,770 Renal: 240,404 Surgical Divisions Cancer: 89,276 Cardiothoracic: 111,546 General: 183,637 Neurosurgery: 27,144 Orthopedics: 665,995 Total Development Cohort: 3,894,235	
Dataset #2: **ICD-10 Re-specification Dataset** **(ICD-10)**	Most of the results presented here relate to the ICD-10 re-specified measure. Those results are based on a full year of admission data from July 1, 2016 – June 30, 2017, with 12 months of history, for the service-line divisions and risk variables. The total number of admissions prior to inclusions and exclusions in this cohort was 10,069,004. More specifically, we constructed an administrative dataset using: 1. An index dataset that contains administrative inpatient hospitalization data, enrollment data, and post-discharge mortality status for Medicare FFS beneficiaries hospitalized from July 1, 2016 – June 30, 2017. 2. A history dataset that includes inpatient hospitalization data on each patient for the 12 months prior to the index admission; this was used for case-mix risk adjustment. 3. Enrollment and mortality status were obtained from the EDB, which contains beneficiary demographic, benefit, coverage, and vital status information. The EDB is also used to obtain hospice enrollment data between July 2016 to June 2017, which identified patients for whom mortality outcome was not a reasonable signal of care quality. Number of Hospitals = 4,692 Patient Descriptive Characteristics: mean age = 77.5; standard deviation = 7.9 The number of index admissions (the cohort) was 6,514,038 following application of the inclusion criteria.	• Section 2b1 Data Element & Measure Score Validity • Section 2b2 Testing of Measure Exclusion • Section 2b3.3a Identification and selection of risk-adjustment variables • Section 2b3.4b Selection of Social Risk Factors • Section 2b3.6 Statistical model discrimination statistics • Section 2b3.7 Risk model calibration statistics • Section 2b4 Meaningful Differences

	The final study cohort was 4,335,530 following the application of the exclusion criteria. The cohort shown by service-line division: Non-Surgical Cancer: 35,143 Cardiac: 538,655 Gastrointestinal: 322,816 Infectious Disease: 562,168 Pulmonary: 485,122 Renal: 332,176 Orthopedic: 133,485 Neurology: 231,629 Other non-surgical: 402,925 Surgical Cardiothoracic: 133,090 General: 206,256 Orthopedics: 677,336 Cancer: 86,062 Neurosurgery: 31,066 Other surgical: 157,601	
Dataset #3: Split-Sample Dataset	For calculating split-sample reliability, we used a dataset that combined the ICD-10 Re-specification Dataset (dataset 2; July 1, 2016 – June 30, 2017), supplemented with claims, enrollment, and hospice data for admissions between October 1, 2015 – June 30, 2016, to create a dataset with 21 months of claims. Total number of measured entities: 4708 Total number of admissions: 7,762,656	• Section 2a2, Split-sample reliability testing • Section 2b3.6 Predictive ability

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.

As detailed below and in section 2b3.4b, we considered a patient-level sociodemographic status (SDS) variable (Medicare-Medicaid dual-eligibility status) and a composite measure (the AHRQ-validated Socioeconomic Status [SES] index score).

We selected social risk factors variables to analyze after reviewing the literature and examining available national data sources.

In selecting variables, our intent was to be responsive to the National Quality Forum (NQF) guidelines for measure developers and the findings of recent work funded by the IMPACT Act [1,2]. Our approach was to examine patient-level indicators both SES that are reliably available for all Medicare beneficiaries and linkable to claims data and to select those that have established validity.

The SES variables that we examined are:

- Dual-eligible status
- AHRQ-validated SES Index score (summarizing the information from the following variables: 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)

Similarly, we recognize that Medicare-Medicaid dual eligibility has limitations as a proxy for patients' income or assets because it does not provide a range of results and is only a dichotomous measure. However, the eligibility threshold for over 65-year-old Medicare patients is valuable, as it considers both income and assets and is consistently applied across states. Additionally, patients’ dual eligibility for Medicare and Medicaid is an indicator whose data are readily available for use. For the dual-eligible variable, there is a body of literature demonstrating differential health care and health outcomes among beneficiaries, indicating that these variables, while not ideal, allow us to examine some of the pathways of interest [1].

Finally, we selected the AHRQ-validated SES Index score because it is a well-validated variable that describes the average SES of people living in defined geographic areas [3]. Its value as a proxy for patient-level information is dependent on having the most granular-level data with respect to communities that patients live in. We used data from the American Community Survey to create AHRQ SES Index scores at the census block group level and then mapped them to 9-digit ZIP codes via vendor software. The patient-level Medicare FFS claims data were then linked to the AHRQ SES Index scores by patients’ ZIP codes. Given the variation in cost of living across the country, we adjusted the median income and median property value components of the AHRQ SES Index by regional price parity values published by the Bureau of Economic Analysis. This provided a better marker of low-SES neighborhoods.

References

1. Department of Health and Human Services, Office of the Assistant Secretary of Planning and Evaluation. Report to Congress: Social Risk factors and Performance Under Medicare’s Value-based Payment Programs. 2016; https://aspe.hhs.gov/pdf-report/report-congress-social-risk-factors-and-performance-under-medicares-value-based-purchasing-programs. Accessed November 10, 2017.

2. National Academies of Sciences, Engineering, and Medicine (NASEM); Accounting for Social Risk Factors in Medicare Payment: Data. Washington DC: National Academies Press; 2016.

3. Bonito A, Bann C, Eicheldinger C, et al. Creation of new race-ethnicity codes and socioeconomic status (SES) indicators for Medicare beneficiaries. Final report, sub-task. 2008; 2.

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 "split sample" 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. We then compared the agreement between the two resulting performance measures across hospitals [1].

We estimated the overall measure score reliability by calculating the intra-class correlation coefficient (ICC) using a split-sample method. This form of measure reliability testing evaluates, on a whole, how reliable measure results are across all facilities. Dataset 3 was used to calculate split-sample reliability. Admissions were randomly and evenly split into the two split samples within each individual hospital (21 months of combined data). For each sample, we fit a hierarchical generalized linear model for each service-line division and then aggregate the results into an overall risk-standardized mortality rate (RSMR). That is, each hospital will have a RSMR in each split sample. The ICC evaluates the agreement between the risk-standardized mortality rates (RSMR) calculated in the two randomly split samples. The ICC estimated was ICC [2, 1], described in Shrout and Fleiss [2], and assessed using conventional standards [3].

We do not provide signal-to-noise reliability for the overall RSRM score because the signal-to-noise calculation should be based on a statistical model [4]; the measure score (RSMR) of the HWM measure is a combined score that is not calculated from a single statistical model.

References

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

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

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

4. Adams J., The reliability of provider profiling: A tutorial. RAND Health, 2009. https://www.rand.org/content/dam/rand/pubs/technical_reports/2009/RAND_TR653.pdf; accessed on January 4, 2019

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

As a metric of agreement, we calculated the ICC [1,2]. To calculate the ICC, we used Dataset 3. In total, 7,762,656 admissions and all hospitals were included in the analysis, using 21 months of data (Note that we only retain hospitals that had have results within both time periods in Dataset 3, or 4,450 hospitals for this analysis). The agreement between the two independent assessments of the RSMR for each hospital was 0.8187, and the adjusted ICC (which estimates the ICC if we had been able to use one full year of data in each split sample) [3,4], is 0.8377 (Table 1). Both demonstrate high reliability, according to conventional standards [1].

Table 1: Claims-only HWM Split-sample Reliability

Statistic	Split-sample reliability (all hospitals)
Number of hospitals	4450
ICC [2,1]	0.8187
Adjusted ICC [2,1]	0.8377

References

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

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

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

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

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 overall split-sample reliability score (0.8377) is interpreted to indicate high reliability.

Our interpretation of these results is based on the standards established by Landis and Koch (1977) [1]:
< 0 – Less than chance agreement;
0 – 0.2 Slight agreement;
0.21 – 0.39 Fair agreement;
0.4 – 0.59 Moderate agreement;
0.6 – 0.79 Substantial agreement;
0.8 – 0.99 Almost Perfect agreement; and
1 Perfect agreement

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

2b1. VALIDITY TESTING

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

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

☒ **Performance measure score**

☒ **Empirical validity testing**

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

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

Empirical Validity Testing of the Measure Score

In order to test the validity of the HWM measure score, we examined whether better performance on the claims-only HWM measure was related to better performance for other relevant structural and outcome measures. However, there are multiple challenges associated with this approach:

1. There are many measures that use a variety of criteria to define a high performing hospital, including: adherence to core processes of care, complications and safety measures, and patient satisfaction.

2. Together with our Technical Workgroup, which consists of nationally recognized experts in measure development, as well as other measurement experts, we have concluded that there is no single recognized and accepted "gold standard" measure that specifically measures factors most relevant to such a broad measure as Hospital-Wide Mortality (HWM). Our approach was to select three separate assessments against which we could compare the measure score with the hypothesis that a trend

toward correlation with these external assessments would support a conclusion of high measure score validity.

After reviewing available measures, we selected the following three to use for validity testing.

1. Nurse-to-bed ratio: Several studies have found that higher levels of nurse staffing are associated with improved patient outcomes and lower mortality rates. [1-4]. We used a nurse-to-bed ratio calculated using two fields from the American Hospital Association's (AHA) annual survey. The AHA surveys all hospitals in the United States and the response rate averages 85–95 percent annually [5], covering about 6,000 hospitals. Staffing is measured as the numbers of full-time and part-time RNs, and LPNs. Within the American Hospital Associations annual survey from 2016, we used the fields "FTEN" and "HOSPBD", which are self- reported fields that are defined in the AHA data dictionary as: number of reported full-time registered nurse and number of hospital beds.

2. Hospital Star Rating mortality group score: CMS's Overall Hospital Star Rating assesses hospitals' overall performance (expressed on *Hospital Compare* graphically, as stars) based on a weighted average of group scores from different domains of quality (mortality, readmissions, safety, patient experience, imaging, effectiveness of care, timeliness of care). The mortality group is comprised of the mortality measures that are publicly reported on hospital compare. The mortality group score is derived from a latent-variable model that identifies an underlying quality trait for that group. For the validity testing presented in this testing form, we used mortality group scores from 4581 Medicare FFS hospitals from July 2018. The full methodology for the Overall Hospital Star Rating can be found at: https://www.qualitynet.org/dcs/ContentServer?c=Page&pagename=QnetPublic%2FPage%2FQnetTier3&cid=1228775957165

3. Overall Hospital Star Rating: CMS's Overall Hospital Star Rating assesses hospitals' overall performance (expressed on *Hospital Compare* graphically, as stars) based on a weighted average of "group scores" from different domains of quality (mortality, readmissions, safety, patient experience, imaging, effectiveness of care, timeliness of care). Each group has within it, measures that are reported on *Hospital Compare*. Group scores for each individual group are derived from latent-variable models that identify an underlying quality trait for each group. Group scores are combined into an overall hospital score using fixed weights; overall hospital scores are then clustered, using k-means clustering, into five groups and are assigned one-to-five stars (the hospital's Star Rating). For the validity testing presented in this testing form, we used hospital's Star Ratings from 4581 Medicare FFS hospitals from July 2018. The full methodology for the Overall Hospital Star Rating can be found at https://www.qualitynet.org/dcs/ContentServer?c=Page&pagename=QnetPublic%2FPage%2FQnetTier3&cid=1228775957165.

We examined the relationship of performance on the claims only measure scores (RSMR) with each of the three external measures of hospital quality. For the external measures, the comparison was against performance within quartiles for nurse-to-bed ratio and mortality group score, or in the case of Star Ratings, to the Star Rating category (1-5 Stars).

We also compared performance on these external measures with categories of performance on the HWM measure by determining "outliers" of performance for the RSMR. Specifically, we identified outliers by estimating an interval estimate (similar to a confidence interval) around each hospital score and identified those facilities that had a 95% interval estimate entirely above or entirely below the national average. We then assigned scores to one of three performance categories: 1) "no different than national average," 2) "better than the national average," or 3) "worse than the national average." – with 95% confidence. Hospitals categorized as outliers ("better" or "worse" than national average) on the HWM measure were identified within the quartiles of performance on the comparator measure (see Figures 1, 2, and 3 in section 2b1.3).

Face Validity as Determined by the TEP

We systematically assessed the face validity of the HWM claims-only measure score as an indicator of quality by confidentially soliciting the TEP members' agreement with the following statement via an online survey following the final TEP meeting: "The risk-standardized hospital mortality rates obtained from the claims-only HWM measure as specified can be used to distinguish between better and worse quality facilities." The survey offered participants response options on a six-point scale (1=Strongly Disagree, 2=Moderately Disagree, 3=Somewhat Disagree, 4=Somewhat Agree, 5= Moderately Agree, and 6=Strongly Agree).

TEP members:

Jeanne Black, PhD, MBA
Manager of Health Policy and Program Evaluation
Cedars-Sinai Health System

John Bott, MBA, MS
Independent Consultant

Roger Dmochowski, MD, MMHC, FACS
Executive Medical Director of Quality, Safety, and Risk
Vanderbilt University Medical Center

Richard Dutton, MD, MBA
Chief Quality Officer
United States Anesthesia Partners

Gaye Hyre
Patient/Family Caregiver Representative
CT State Innovation Model for Healthcare Equity and Access Council Member

Irene Katzan, MD, MS
Director, Neurological Institute for Outcomes Research and Evaluation
Cleveland Clinic

Brenda Matti-Orozco, MD, FACP
Chief of Division of General Internal Medicine and Palliative Medicine
Hospice Medical Director
Morristown Medical Center and Atlantic Home Care & Hospice

Michelle Beck
Consumer
University of Maryland
Upper Chesapeake Medical Center

Use of 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" [6,7].

References

1. Aiken LH, Clarke SF, Sloane DM, Sochalski J, Silber JH. Hospital Nurse Staffing and Patient Mortality, Nurse Burnout, and Job Dissatisfaction. Journal of the American Medical Association. 2002;288(16):1987–93.

2. Griffiths P, Ball J, Murrells T, et al. Registered nurse, healthcare support worker, medical staffing levels and mortality in English hospital trusts: a cross-sectional study. BMJ Open 2016;6:e008751. doi: 10.1136/bmjopen-2015-008751

3. Needleman J, Buerhaus PI, Mattke S, Stewart M, Zelevinsky K. Nurse-Staffing Levels and The Quality of Care in Hospitals. New England Journal of Medicine. 2002;346:1719–22.

4. Needleman J., Buerhaus P., Pankratz V.S., Leibson C.L., Stevens S.R., Harris M. Nurse staffing and inpatient hospital mortality. N. Engl. J. Med. 2011;364:1037–1045. doi: 10.1056/NEJMsa1001025.

5. American Hospital Association (AHA) The AHA Annual Survey Database Fiscal Year 1997 Documentation. Chicago: Health Forum; 1999.

6. 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. 2006; 113(3):456-462.

7. National Quality Forum. National voluntary consensus standards for patient outcomes, first report for phases 1 and 2: A consensus report. Available at: http://www.qualityforum.org/projects/Patient_Outcome_Measures_Phases1-2.aspx. Accessed January 6, 2019.

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

Empiric Validity Testing

To examine the external validity of the HWM measure results, we explored the relationship of performance on the HWM claims-only measure scores (overall risk-adjusted HWM rate or RSMR) with each of the three external measures of hospital quality: nurse-to-bed ratio, mortality group score of Star Rating, and Overall Star Rating.

For each external measure of quality, the comparison showed a trend toward better performance on the HWM measure with better performance on the comparator measure. For example, in Figure 1, when comparing the claims-only HWM measure to the nurse to bed ratio, as the number of nurses per bed increases (more nurses in the hospital) across quartiles of nurse-to-bed ratio (from left to right on the graph), the median overall HWM mortality rate is lower (better). Likewise, in Figure 2, better performance on the HWM measure is associated with better Star Rating mortality group scores across quartiles of mortality group score performance. Finally, in Figure 3, we show that HWM performance improves across the Star Rating category in the expected direction: HWM scores are better (lower) as the Star Rating category improves (increases from 1, to 5 Stars).

Within the graphs, we also overlay "better" and "worse" outliers (95% confidence interval, as described in 2b1.2) on the HWM measure, with performance on the external measure. The overlay results are consistent with the trend toward better performance on the HWM measure with better performance on the quality measure; there are more high outliers (shown in total as "better" at the bottom of the graph, and as blue squares in the graph) with higher performance for each comparator measure (moving left to right on the graphs below); there are also fewer "worse" outliers (shown in total as "worse" at the bottom of the graph, and as red triangles in the graph). The inverse is also observed: fewer "better" outliers and more "worse" outliers are present in quartiles of worse performance on the comparator measure.

For example, in Figure 1 below, which compares RSMR to nurse-to-bed ratio, there are:

- 19 HWM "better than national average" outliers in the third quartile (and 5 "worse" outliers)
- 49 HWM "better" outliers in the fourth quartile (and zero "worse" outliers)

In addition, in Figure 2, which compares RSMR to mortality group score, there are:

- 96 HWM "better than national average" outliers (and zero "worse" outlies) in the fourth (best performing) quartile
- 13 "worse than national average" (and zero "better") outliers in the first (worst performing) quartile.

A similar relationship can be seen in Figure 3, in comparison to the Overall Hospital Star Rating.

Figure 1: RSMR Relationship to Nurse-to-Bed Ratio

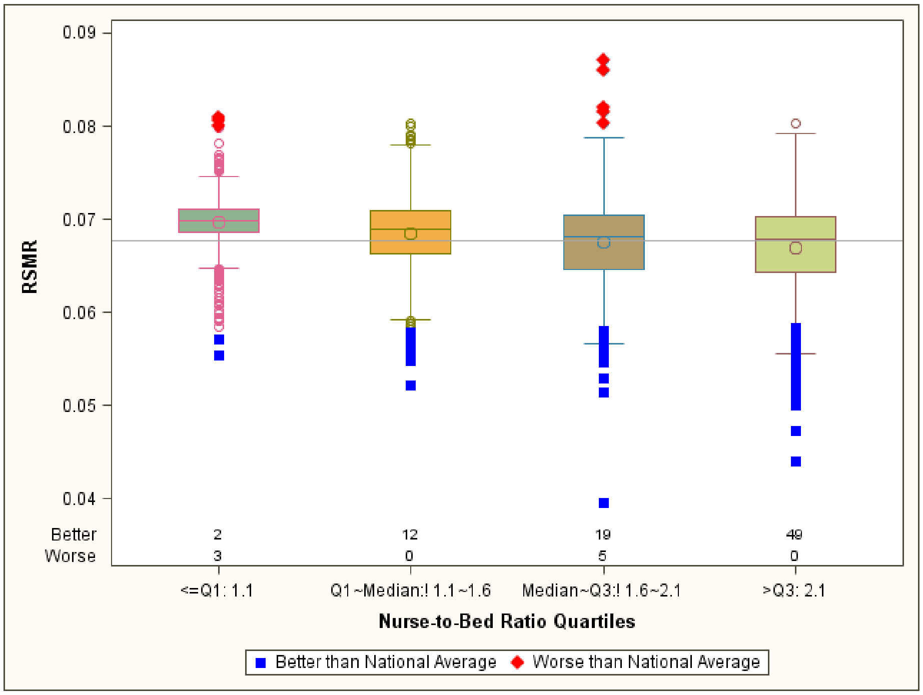


Figure 2: RSMR Relationship to Star Ratings Mortality Group Score

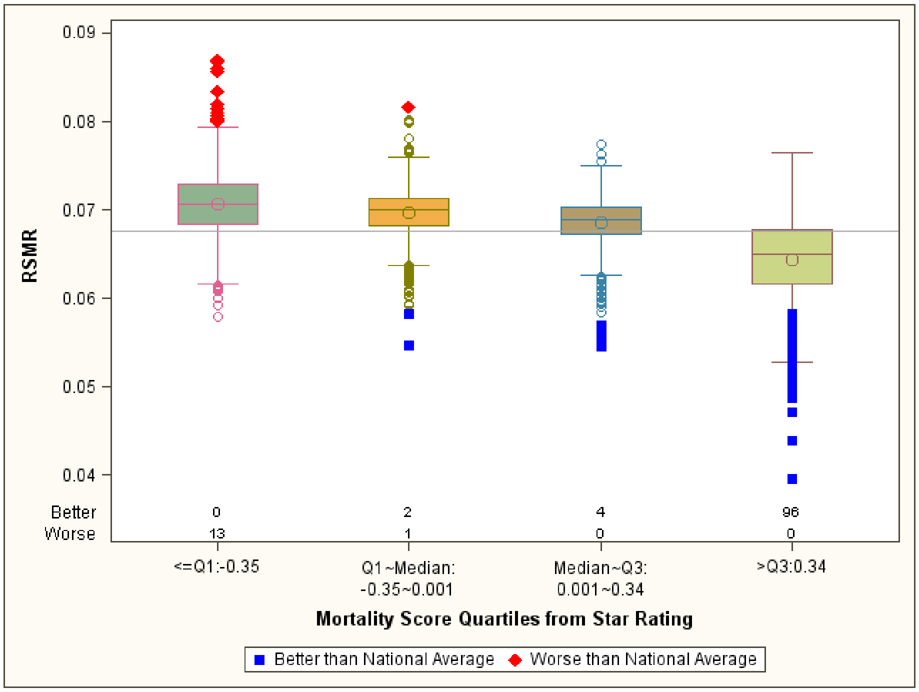
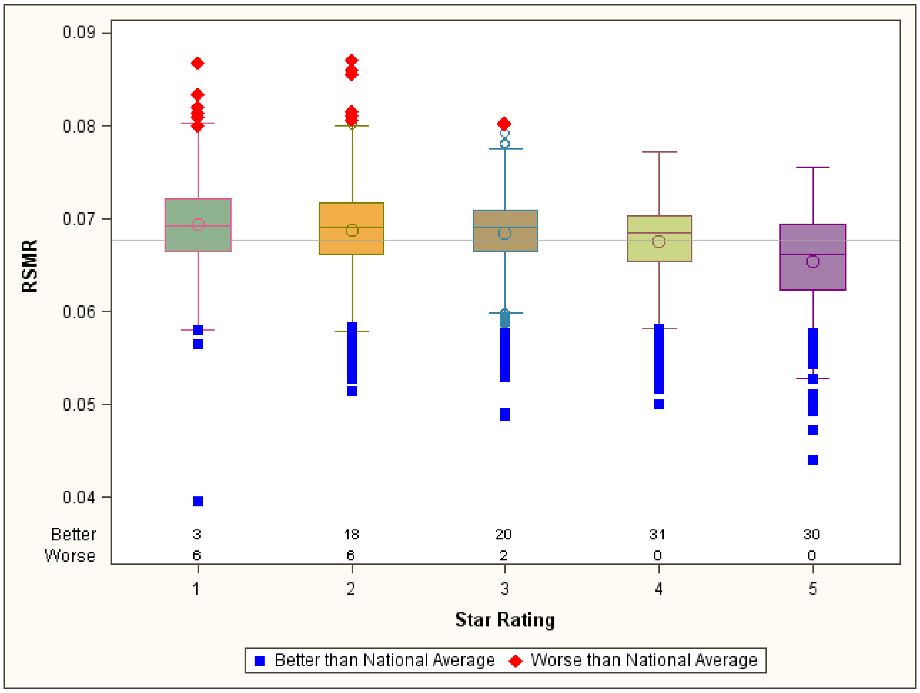


Figure 3: RSMR Relationship to Overall Star Rating



Validity as assessed by the TEP

Validity was assessed by the TEP using a post-meeting survey. The TEP provided input on the cohort, risk model, and outcome to strengthen the measure and supported the final measure with high agreement. A total of 6 TEP members completed the face validity survey. Of the 6 respondents, 5 respondents (83%) indicated that they somewhat, moderately, or strongly agreed, and 1 moderately disagreed, with the following statement: "The risk-standardized hospital mortality rates obtained from the claims-only HWM measure, as specified, can be used to distinguish between better and worse quality facilities."

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?*)

The results above show that the HWM measure agrees with external measures of quality. There is a trend in the expected direction that overlaps and matches all three external measures of quality, which provides external support for measure score validity.

In addition, the "better" and "worse" outliers for HWM align with performance on each external measure of quality. In other words, there are a higher proportion of HWM "better" performers in higher performing categories for each measure. The reverse is also true for two of the three comparator measures: there are more "worse" HWM outliers as you move left on the graph (toward worse performance on the external measure).

Please note that, based on our discussions with our Technical Workgroup and with other experts, we concluded that there is no single analysis that is sufficient to validate the measure because there is no gold standard exists for the validation of a hospital-wide quality measure. With this limitation in mind, we used the three empiric external analyses to demonstrate a trend of validity using different metrics.

In addition, the Overall Star Rating includes quality measures that are much broader than the HWM Measure, such as patient experience. This is consistent with the stronger relationship that can be seen between HWM and the Star Rating mortality measure group score (Figure 2).

Survey results from the TEP indicate relatively high agreement (83%) regarding the face-validity of the claims-only HWM measure.

____

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. Rationales for the exclusions are detailed in section S.8 of the Submission/Intent to Submit form (Denominator Exclusions).

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

Applying our inclusion criteria (See section S.6 of the Intent to Submit/Submission form) resulted in an initial cohort of 6,514,038. We then applied the following exclusion criteria (see the Intent to Submit Form, sections S.8 and S.9, for exclusion rationale) with the following number and percent of excluded admissions (referenced as a percent of the initial cohort):

Discharged against medical advice: 28,739 (0.44%)
Inconsistent or unknown vital status or other unreliable data: 87 (0.0013%)
Primarily treated for crush injury, spinal cord injury, intracranial injury, or burns: 61,519 (0.94%)
Fewer than 100 admissions in a CCS within a division: 13,597 (0.21%)

Given the few cases affected, we did not examine the distribution of admissions across hospitals or the effect of the exclusion on measure scores.

Note that our final cohort is 4.3 million, due to a processing step that requires the random selection of a single admission for patients with multiple admissions. Random selection ensures that providers are not penalized for a 'last' admission during the measurement period; selecting the last admission would not be as accurate a reflection of the risk of death as random selection, as the last admission is inherently associated with a higher mortality risk. Random selection is also used in CMS's condition-specific mortality measures. Note that random selection reduces the number of admissions, but does not exclude any patients from the measure.

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

The exclusions for this measure are narrowly targeted. The largest exclusion (0.94% of admissions) is clinically defined and reflects the inability to adequately risk adjust given that trauma is unevenly distributed across hospitals. In total, exclusions remove a small number (about 1.4%) of admissions.

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 21 risk factors**

☐ **Stratification by** Click here to enter number of categories **risk categories**

☐ **Other,** Click here to enter description

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

<u>**Risk Model**</u>:
The method estimates a separate hierarchical logistic regression model for each service-line division. In order to obtain the variance and interval estimates, the hierarchical model is fit under the Bayesian framework along with the Markov Chain Monte Carlo (MCMC) technique. Details of the risk model, including equations, can be found in the data dictionary, tab "HWM_Statistical Approach", as well as in the technical report, in section 4.6.1, which is attached to this application.

<u>**Risk Factors**</u>:
The goal of risk adjustment is to account for differences across hospitals in patient demographic and clinical characteristics that might be related to the outcome but are unrelated to quality of care. Risk adjustment for this measure was complicated by the fact that it includes many different discharge condition categories, as well as patients undergoing surgical procedures. Therefore, this measure adjusts for both case mix differences (clinical status of the patient on admission, accounted for by adjusting for comorbidities and diagnoses present on admission) and service mix differences (the types of conditions/procedures cared for by the hospital, accounted for by adjusting for the discharge condition category).

(1) Case-mix adjustment: Comorbid Risk Factors (defined by CMS's Condition Categories – CCs):

- age
- pneumonia (CC 114-116)
- dialysis or severed chronic kidney disease (CC 134, 136, 137)
- acute or unspecified renal failure (CC 135, 140)
- poisonings and allergic and inflammatory reactions (CC 175)
- minor symptoms, signs, findings (CC 179)
- protein-calorie malnutrition (CC 21)
- disorders of fluid/electrolyte/acid-base balance (CC 24)
- disorders of lipoid metabolism (CC 25)
- liver failure (CC 27, 30)
- other GI disorders (CC 34, 35, 37, 38)
- other musculoskeletal and connective tissue disorders (CC 44, 45)
- hematologic or immunity disorders (CC 46-48)
- dementia and other nonpsychotic organic brain syndromes (CC 51-53)
- other infectious diseases (CC 7)
- metastatic & severe cancers (CC 8, 9)
- coma/brain compression/ anoxic injury and severe head injury (CC 80, 166)
- respiratory failure, respirator dependence, shock (CC 82-84)
- congestive heart failure (CC 85)
- hypertension and hypertensive heart disease (CC 94, 95)

Primary and secondary diagnoses codes identified as potential complication of care (see tab HWM_Complications of the data dictionary) with an associated "present on admission" code are kept in the risk model; any potential complication of care without an associated “present on admission” code are removed from the risk model under the assumption that it represented a complication of care.

Comorbid risk variables are the same for each of the 15 divisions, but coefficients for each comorbid risk factor vary by division; see tab HWM_Risk_Var_ParEst for the coefficients for each comorbid risk factor, by division.

(2) Risk factors: Service-Line risk adjustment:

As described in section S.7 of the submission form, for the cohort we use the AHRQ CCS grouper to group all ICD-10 principal discharge diagnoses into clinically coherent categories.

For risk adjustment, as described in section 2b3.3a., for all AHRQ principal discharge diagnosis code CCSs, we include a discharge diagnosis-specific indicator in the model. This ensures that the principal discharge diagnosis for each patient is also included in the risk model, in addition to the 21 variables described above.

Discharge diagnosis categories differ in their baseline mortality risks and hospitals will differ in their relative distribution of these discharge diagnosis categories (service mix) within each division. Therefore, adjusting for principal discharge diagnosis categories levels the playing field across hospitals with different service mixes.

See the data dictionary for the CCSs (tabs HWM Non-SurgCohortDiv CCS and HWM SurgicalCohortDiv CCS) that comprise each of the divisions in this measure, and the parameter estimates for the different CCS categories within each of the 15 divisions (HWM_Risk_Var_ParEst). Also see tab “HWM_CCS_Modifications” for the 20 CCSs that were modified by our clinical consensus process, defined in section 2b3.3a.

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

Not applicable. This measure is risk adjusted.

2b3.3a. Describe the conceptual/clinical <u>and</u> 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?

<u>Selecting Risk Variables</u>

<u>Candidate Comorbid Risk Variables</u>

Our goal is to develop parsimonious models that include clinically relevant variables strongly associated with the risk of mortality in the 30 days following an index admission. For candidate variable selection, using the development sample we started with the CMS Condition Categories (CC)s grouper, used in previous CMS risk-standardized outcome measures, to group ICD-9 codes into comorbid risk adjustment variables.

To select candidate variables, a team of clinicians reviewed all CMS-CCs and combined some of these CMS-CCs into clinically coherent groups to ensure adequate case volume. Any combined CMS-CCs were combined using both clinical coherence and consistent direction of mortality risk prediction across the CMS-CC groups in the majority of the 15 divisions.

Potential Complications of Care During Hospitalization

Complications occurring during hospitalization are not comorbid illnesses and do not reflect the health status of patients upon presentation. In addition, they likely reflect hospital quality of care, and, for these reasons, should **not** be used for risk adjustment. Although adverse events occurring during hospitalization may increase the risk of mortality, including them as risk factors in a risk-adjusted model could lessen the measure's ability to characterize the quality of care delivered by hospitals. We have previously reviewed every CMS-CC and identified those which, if they were to occur only during the index hospitalization, are more likely than not to represent potential complications rather than pre-existing comorbidities. For example: fluid, electrolyte, or base disorders; sepsis; and acute liver failure are all examples of CMS-CCs that could potentially be complications of care.

For the claims-only HWM measure, we took a two-step approach to identifying complications of care. First, we searched the secondary diagnosis codes in the index admission claim for all patients in the measure and identified the presence of any ICD-9 code associated with a CMS-CC. If these codes appeared only in the index admission claim, we flagged them because they are potential complications of care. Next, we determined if these potential complications of care were associated with a "present on admission" code. Any potential complication of care with an associated "present on admission" code was kept in the risk model; any potential complication of care without an associated "present on admission" code was removed under the assumption that it represented a complication of care. In this way, we supplemented the existing approach to identifying potential complications of care used in CMS's publicly reported mortality measures by incorporating "present on admission" codes. Our analyses demonstrate that a majority of hospitals currently use "present on admission" codes across a majority of conditions. Therefore, we felt that a combined approach to excluding complications of care from the risk model that used both the existing methodology and "present of admission" codes allow the measure to capture as many clinically appropriate risk variables as possible while simultaneously removing complications of care from the risk model.

Final Comorbid Risk Variable Selection

To inform variable selection, we used the development sample to create 500 bootstrap samples for each of the service-line divisions (this analysis was performed prior to removing the divisions Other Non-Surgical Conditions and Other Surgical Procedures; therefore, this analysis was completed on 15 divisions). For each sample, we ran a standard logistic regression model that included all candidate variables. The results were summarized to show the percentage of times that each of the candidate variables was significantly associated with 30-day mortality (at the p<=0.05 level) in the 500 bootstrap samples (for example, 90% would mean that the candidate variable was significant at p<=0.05 in 90% of the bootstrap samples). We also assessed the direction and magnitude of the regression coefficients.

We found that models containing all risk factors performed similarly to models containing a more limited set of "significant" risk factors, described below. We therefore used a fixed, common set of comorbidity variables in all of our models for simplicity and ease of implementation and analysis. We describe below the steps for variable selection.

a. The CORE Project Team reviewed the bootstrapping results and decided to provisionally examine risk adjustment variables at or above a 90% cutoff in one of the 15 service-line division models (in other words, retain variables that were significant at the p<=0.05 level in at least 90% of the bootstrap samples for each division). We chose the 90% cutoff because this threshold has been used across other measures and produced a model with adequate discrimination.

b. In order to develop a statistically robust and parsimonious set of comorbid risk variables, we then chose to limit the variables to those that met a 90% threshold in at least 13/15 divisions. This step resulted in the retention of 20 risk factors, including age and 19 comorbid risk variables. This resulted in C-statistics that did not change by more than 0.02 in any of the 15 divisions compared to models that contained all possible risk variables.

Service-Line Adjustment

As described in section S.7 of the intent to submit form, we use the AHRQ CCS grouper to group all ICD-10 principal discharge diagnoses into clinically coherent categories (categories have been somewhat modified as described below). For all AHRQ principal discharge diagnosis code CCSs with sufficient volume (CCSs with fewer than 100 admissions are excluded), we also included a discharge diagnosis-specific indicator in the model. This ensures that the principal discharge diagnosis for each patient is also included in the risk model, in addition to the 20 comorbid risk variables described above.

Rationale: Discharge diagnosis categories differ in their baseline mortality risks and hospitals will differ in their relative distribution of these discharge diagnosis categories (service mix) within each division. Therefore, adjusting for principal discharge diagnosis categories levels the playing field across hospitals with different service mixes. See the data dictionary for the CCSs (tabs HWM Non-SurgCohortDiv CCS and HWM SurgicalCohortDiv CCS) that comprise each of the divisions in this measure and HWM_Risk_Var_ParEst for the parameter estimates for the CCS categories for each division.

<u>CCS modifications</u>: Note that in addition to using the AHRQ CCS grouper to define the CCS categories in each division (see section S.7 of the submission form), we made two types of modifications: (1) We modified selected CCS highly heterogenous CCS categories to create more homogenous CCS risk variable groups, and so increased the face validity of risk model, described below, and (2) we combined low-mortality CCSs (those with mortality rate of 1% or lower), also described below.

Heterogenous CCSs: In parallel with our approach during measure development in ICD-9 (see section 4.5.3 of the attached technical report) and in response to feedback from our TEP and Technical Workgroup, we addressed heterogeneity within specific AHRQ CCS groups where the risk of mortality varied significantly across the different ICD-10 diagnoses within the CCS. As explained in detail in the technical report, we calculated the correlation between mortality rates grouped by principal discharge diagnosis ICD-10 code within each CCS. We identified any CCS with an intra-class correlation (ICC) score >0.05 as having high heterogeneity. (The ICC is used in this context to identify heterogeneity of mortality risk across ICD-10 codes within the ICC. The value 0.05, or 5%, is a conventional threshold for accounting for between group heterogeneity.) To address the heterogeneity, three clinicians independently, and through consensus, modified the highly heterogeneous CCSs using clinically informed recategorizations, by either splitting the CCSs into more than one CCS, moving ICD-10 codes to more clinically coherent CCSs, or removing from inclusion ICD-10 codes where quality of care less likely impacts survival, and/or where there were a small number of patients. During ICD-10 re-specification, we identified 44 highly heterogeneous CCSs and made modifications to 20 of them, as described in the data dictionary, tab "HWM_CCS_Modifications."

Low-mortality CCSs: During initial measure development, the patient-level risk models for two divisions (the "Other" surgical and non-surgical divisions) did not converge due to the large number of CCS category codes in these divisions, and due to low mortality rates associated with some of the CCSs in these divisions (which are used for service-line risk adjustment). However, the TEP and Patient and Family Caregiver Workgroup had a strong interest in retaining these admissions (more than half a million admissions) in the measure. To address this issue, within each division, CCSs with low mortality rates (those less than or equal to 1%) are combined into one independent group, which reduces the total number of risk variables (CCS category codes) in the model.

Social Risk Factors for Disparities Analyses

We selected variables representing social risk factors based on a review of literature, conceptual pathways, and feasibility. In section 1.8, we describe the variables available in Medicare claims data that we considered and analyzed based on this review. Below, we describe the pathways by which social risk factors may influence risk of the outcome.

Our conceptualization of the pathways by which patients' social risk factors affect the outcome is informed by the literature cited below and IMPACT Act–funded work by the National Academy of Science, Engineering and Medicine (NASEM) [5] and the Department of Health and Human Services Assistant Secretary for Policy and Evaluation (ASPE) [7].

Causal Pathways for Social Risk Variable Selection

There is a large body of literature linking various SES factors to worse health status and higher mortality over a lifetime [2, 6, 14, 22]. 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, [4, 10, 13, 17, 18]). 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 [8]. 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 [1]. 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 [11, 12].

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

1. 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.

2. Brodish P.H., Hakes J.K. “Quantifying the individual-level association between income and mortality risk in the United States using the National Longitudinal Mortality Study.” Soc. Sci. Med., 170 (2016), pp. 180-187, 10.1016.

3. 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.

4. 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

5. Committee on Accounting for Socioeconomic Status in Medicare Payment Programs; Board on Population Health and Public Health Practice; Board on Health Care Services; Institute of Medicine; National Academies of Sciences, Engineering, and Medicine. Accounting for Social Risk Factors in Medicare Payment: Identifying Social Risk Factors. Washington (DC): National Academies Press (US); 2016 Jan 12. (https://www.ncbi.nlm.nih.gov/books/NBK338754/doi:10.17226/21858)

6. Demakakos P, Biddulph JP, Bobak M, Marmot MG (2016a) Wealth and mortality at older ages: a prospective cohort study. J Epidemiol Community Health 70:346–353.

7. Department of Health and Human Services, Office of the Assistant Secretary of Planning and Evaluation. Report to Congress: Social Risk Factors and Performance under Medicare’s Value-based Payment Programs. December 21, 2016. (https://aspe.hhs.gov/pdf-report/report-congress-social-risk-factors-and-performance-under-medicares-value-based-purchasing-programs).

8. 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.

9. 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.

10. 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

11. 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.

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

13. 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

14. Kim D. The associations between US state and local social spending, income inequality, and individual all-cause and cause-specific mortality: The National Longitudinal Mortality Study. Prev. Med. 2015;84:62–68. doi: 10.1016/j.ypmed.2015.11.013.

15. 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.

16. 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.

17. 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

18. 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

19. 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.

20. 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.

21. Shahian DM, Iezzoni LI, Meyer GS, Kirle L, Normand SL. Hospital-wide mortality as a quality metric: conceptual and methodological challenges. *American journal of medical quality: the official journal of the American College of Medical Quality.* 2012;27(2):112-123.

22. van Oeffelen AA, Agyemang C, Bots ML, Stronks K, Koopman C, van Rossem L, Vaartjes I. The relation between socioeconomic status and short-term mortality after acute myocardial infarction persists in the elderly: results from a nationwide study. Eur J Epidemiol. 2012 Aug;27(8):605-13. doi: 10.1007/s10654-012-9700-z. Epub 2012 Jun 5.

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?

Tables showing the final variables in the 15 models with associated odds ratios (OR) and confidence intervals (CI) are in tab HWM_Risk_Var_ParEst of the HWM claims-only data dictionary.

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.**

To examine the impact of social risk factors on the measure calculation, we evaluated two indicators of social risk: Medicaid dual-eligibility and the AHRQ SES index. Our goal for these analyses were two-fold: 1) to examine whether these factors were associated with increased risk in the outcome after adjusting for other risk factors and 2) to evaluate the impact of social risk factors on risk-adjusted HWM rates (RSMRs).

<u>Analysis #1. Distribution of social risk factors across hospitals</u>: To examine how the proportion of patients with each social risk factor varied across hospitals.

The prevalence of social risk factors in the HWM cohort varies across measured entities. The median percentage of dual-eligible patients is 14.6% (interquartile range [IQR]: 9.2% – 22.8%). The median percentage of low-SES patients using the AHRQ SES Index score is 17.0% (IQR: 7.1% – 34.2%).

Analysis #2. Patient-level observed mortality rates for patients with social risk factors

To evaluate the association of these risk factors with the outcome (univariate model), we first quantified the overall observed rate by each social risk factor group (dual-eligible: yes vs. no, AHRQ SES Index: lowest quartile of SES Index vs. all others), as well as the division-level observed rates by each social risk factor.

The overall outcome rate (observed mortality) for patients with dual-eligible status is significantly higher than the outcome rate for patients who are not dual eligible (10.1% vs. 6.3%, p <.0001). The outcome rate for patients with low SES is significantly higher (though the difference not as large as dual-eligible status) than the outcome rate for patients that are not in the lowest SES quartile (7.5% vs. 6%, p<.0001).

We further examined this relationship at the division level, for both dual eligible status (Table 4), and for low SES status (Table 5). At the division level the observed mortality rates for certain divisions (such as infectious disease and surgical, cardiothoracic) are higher for patients with dual eligible status compared to all other patients (Table 4). The differences for patients with low SES status are not as pronounced (Table 5) as those for patients with dual eligible status (Table 4).

Table 4. Division-Level Observed Mortality Stratified by Dual Eligibility status

Division		Non-Dual Eligible		Dual-Eligible	
		Frequency	Observed Mortality Rate (%)	Frequency	Observed Mortality Rate (%)
Non-Surgical	Cancer	30516	14.53	4281	15.42
	Cardiac	467537	6.20	64496	8.67
	Gastrointestinal	272578	4.63	46388	6.62
	Infectious Disease	433904	11.99	121712	17.18
	Other Conditions	331915	5.42	66400	6.68
	Neurology	193960	7.74	34988	9.59
	Orthopedic	113909	4.83	17965	5.92
	Pulmonary	388148	9.18	91183	11.46
	Renal	267963	8.70	60150	9.43
Surgical	Cardiothoracic	121567	5.98	10083	11.64
	Cancer	79180	2.21	5918	3.87
	General	179643	6.07	24310	10.68
	Other Procedures	138808	3.78	17034	6.88
	Neurosurgery	29037	2.97	1745	4.53
	Orthopedics	624641	1.33	45198	3.64

Table 5. Division-Level Observed Mortality Stratified by AHRQ SES status

Division		Not Low SES Status		Low SES Status	
		Frequency	Observed Mortality Rate (%)	Frequency	Observed Mortality Rate (%)
Non-surgical	Cancer	28204	14.31	6749	15.38
	Cardiac	430179	6.38	105861	6.77
	Gastrointestinal	256211	4.79	65064	5.22
	Infectious Disease	443428	12.81	115986	13.89
	Other Conditions	317870	5.55	83063	5.72
	Neurology	184106	8.00	46333	7.85
	Orthopedic	110864	4.93	22108	4.97
	Pulmonary	376388	9.61	106391	9.29
	Renal	254313	8.92	76252	8.21
Surgical	Cardiothoracic	110740	5.99	21772	8.42
	Cancer	72034	2.14	13608	3.28
	General	166277	6.25	39002	8.01
	Other Procedures	127647	3.90	29206	4.92
	Neurosurgery	26569	2.89	4364	3.96
	Orthopedics	581471	1.40	93071	1.91

Analysis #3: Strength and significance of each of the social risk factors in the context of a multivariable model for each division.

We examined the strength and significance of the SES variables in the context of a bivariate model compared with a multivariable model. Consistent with the above findings, when we include these variables in a multivariate model that includes all of the claims-based clinical variables, the odds ratios for both the dual eligible and AHRQ SES variables in the multivariate model are almost always lower than the odds ratio for the bivariate association (middle column of figures) (Figure 4 and Figure 5). This indicates that comorbidity variables that are already in the risk model are attenuating the odds ratios for the social risk factor variables.

For example, in Figure 4, for the dual eligibility risk factor, in the bivariate model (middle column), the odds ratios for this social risk factor are statistically greater than 1.0 in all but one division. However, for each division, including the dual eligibility risk variable in a multivariate model together with all of the claims-based clinical variables (the far right-hand column) attenuates the odds ratios compared with the bivariate model (which only contains the social risk variable). More specifically, in the surgical Orthopedic (“S: Orthopedics”) division, the odds ratio for the bivariate association between dual eligibility and mortality is significantly greater than 1 (middle column), but in a multivariate model, the odds ratios for the outcome for patients with that risk factor are attenuated, and are now no longer significantly different from 1 (far right column). As noted above, this indicates that the comorbid risk variables that are already in the model (in the multivariate view) are capturing the risk associated with the outcome seen in the bivariate analysis (with the social risk factor alone). This means that the dual eligible variable in a multivariate model would not play a significant role in the model (the coefficients/odds ratios are not different from 1). This is true for all of the divisions shown below in Figure 4, except for the surgical neurosurgery division (where the presence of the dual eligible variables is actually protective).

For the AHRQ SES risk factor however (Figure 5), for most of the non-surgical divisions (shown on the lower half of the Figure, labeled "NS"), even the bivariate relationship is not significantly different than one (results span the red line at 1.0), and the AHRQ low SES variable in a multivariate model has little effect on the odds ratio for the outcome for those divisions.

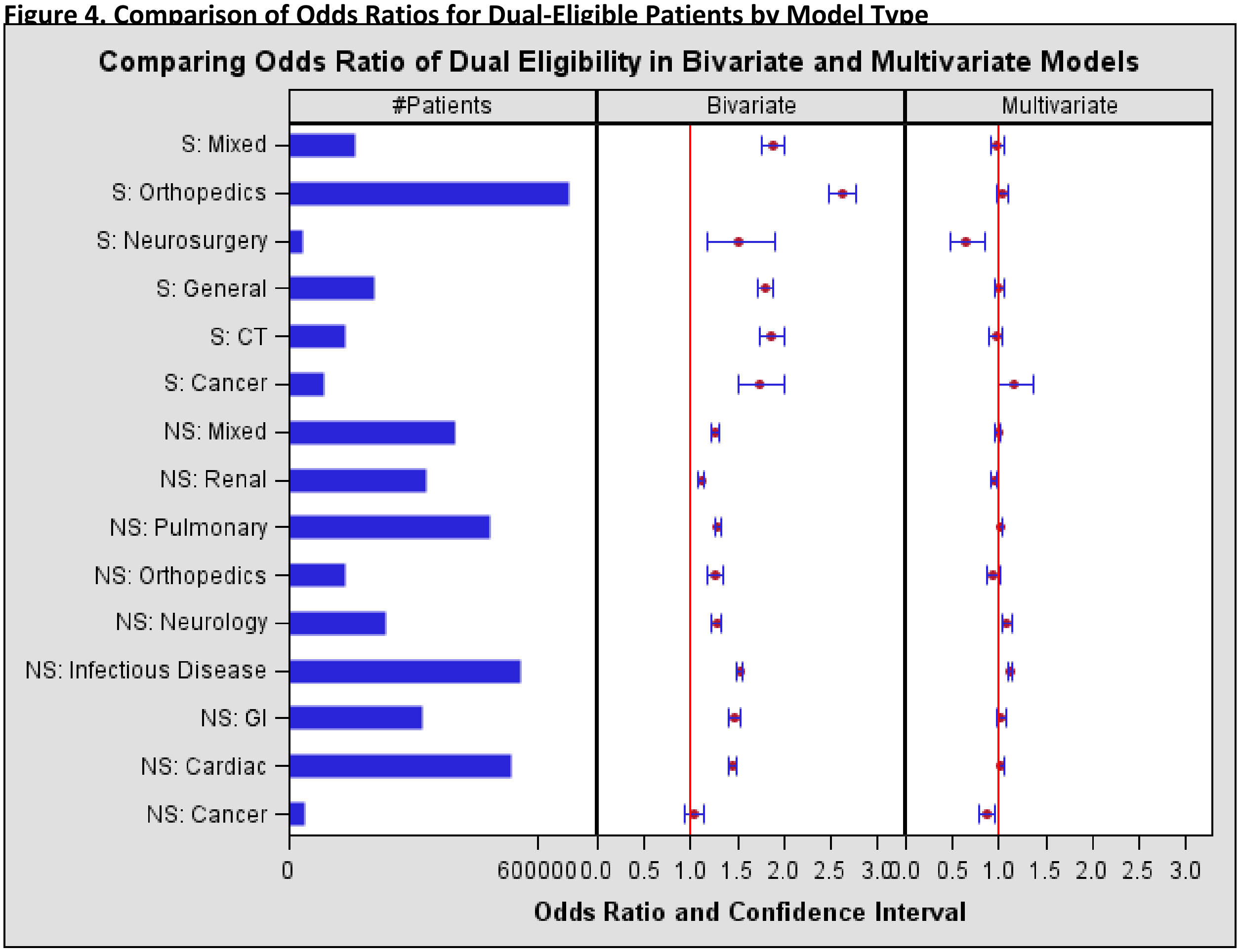
For surgical divisions, the bivariate relationship between the AHRQ SES indicator and the outcome is significant (Figure 5, middle column). However, further analysis shows that adding the risk variable into the multivariate model again attenuates the effect size for most divisions (the odds ratios for most division are close to 1 in the multivariate model), with the exception of the surgical cancer division.

We then separated the hospital- and patient-level effects for the surgical cancer division (see data dictionary tab "HWM_Cancer"), and compared it with two clinical comorbid risk variables (CMS CC 24:Disorders of Fluid/Electrolyte/Acid-Base Balance; and CC 51 Dementia With Complications), and found that for the cancer division:

- There is a small hospital-level effect and no patient-level effect for the dual-eligible variable.
- There is a meaningful hospital-level effect for the low AHRQ SES variable, which is larger than the hospital-level effect of the comparator comorbid risk variables (CC 24; CC 51), but of a similar magnitude to the patient-level effect for this variable.

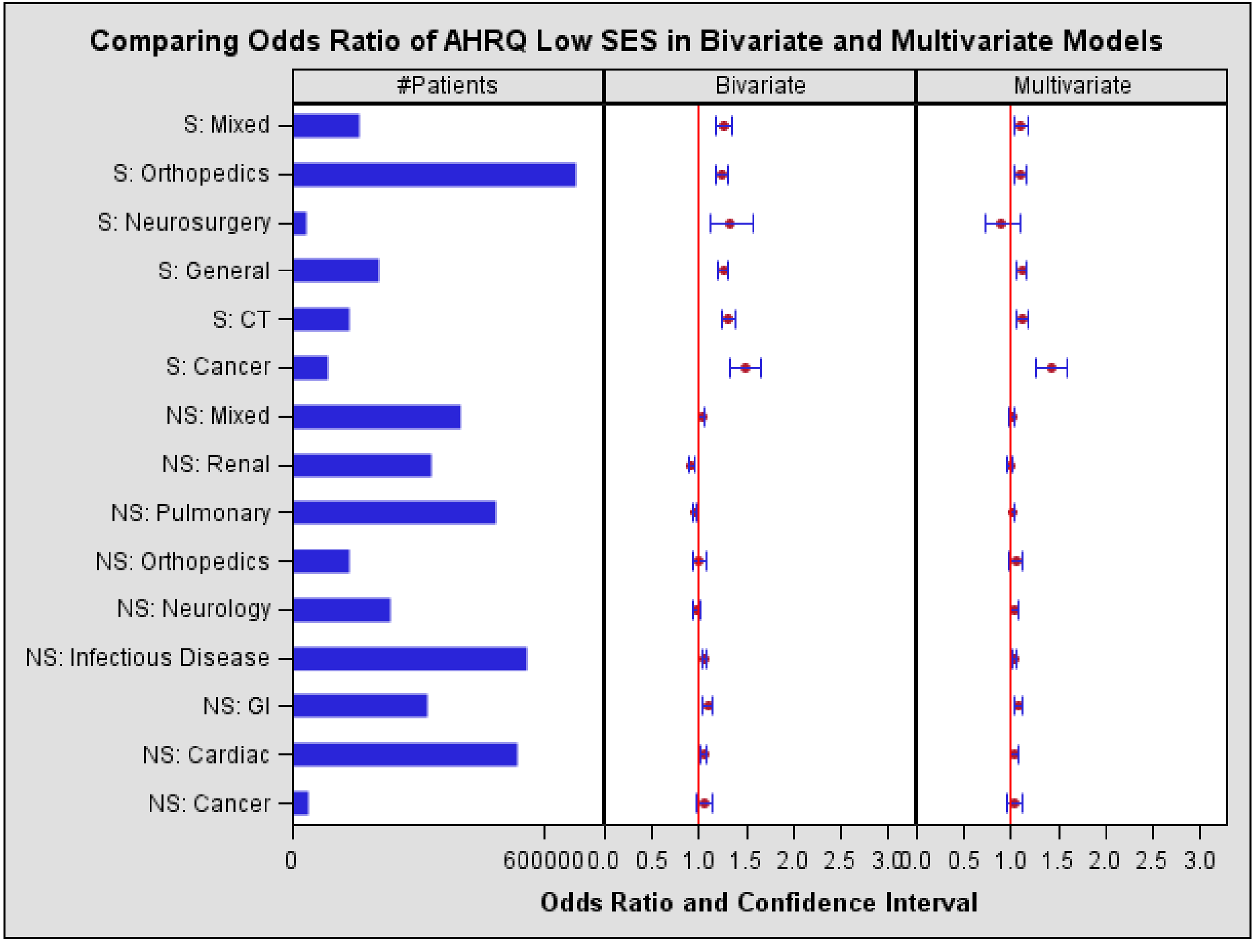
Note that the surgical cancer division is one of the smallest (in terms of admissions) in the measure, representing just 2.0% of admissions in the measure (about 86,000 admissions out of 4.3 million total admissions). This division, therefore, contributes less to the overall HWM score compared to the other divisions.

Figure 4. Comparison of Odds Ratios for Dual-Eligible Patients by Model Type



S = Surgical; S: Mixed = Surgical Other; S: CT = Cardiothoracic surgery; NS = Non-surgical; NS: Mixed=Non-surgical Other

Figure 5. Comparison of Odds Ratios for AHRQ Low SES Patients by Model Type



S = Surgical; S: Mixed = Surgical Other; S: CT = Cardiothoracic surgery; NS = Non-surgical; NS: Mixed=Non-surgical Other

Analysis #5: To understand the effect of each risk factor in the performance and predictive ability of each of the 15 models, we compared the C-statistics for each model with and without the addition of each of the social risk factors. The results shown below in Table 6 indicate that entering these (dual eligible, and low AHRQ SES index) variables into the risk-adjustment model does not improve model performance (c-statistics remained unchanged).

Table 6. Division-Level C-Statistics, with and without Social Risk Adjustment

Division		C-statistic without Social Risk in the Model	C-statistic with Dual-Eligible in the Model	C-statistic with AHRQ SES in the Model
Non-surgical	Cancer	0.78	0.78	0.78
	Cardiac	0.84	0.84	0.84
	Gastrointestinal	0.84	0.84	0.84
	Infectious Disease	0.84	0.84	0.84
	Other Conditions	0.82	0.82	0.82
	Neurology	0.83	0.83	0.83
	Orthopedic	0.82	0.82	0.82
	Pulmonary	0.81	0.81	0.81
	Renal	0.78	0.78	0.78
Surgical	Cardiothoracic	0.83	0.83	0.83
	Cancer	0.85	0.86	0.86
	General	0.87	0.87	0.87
	Other Procedures	0.87	0.87	0.87
	Neurosurgery	0.92	0.92	0.92
	Orthopedics	0.91	0.91	0.91

Analysis #6:

To evaluate the impact of social risk factors on the measure score, we compared RSMRs calculated with and without each social risk factor included in the model. For these analyses we calculated the RSMR difference for each hospital (RSMR with the social risk variable minus RSMR without the social risk variable), and calculated Pearson correlation coefficients for the paired scores. We also show scatter plots for these same analyses.

The results show that entering either of these variables into the risk-adjustment model did not substantially change hospital-level measure scores (RSMRs). The median change in the differences between RSMRs for both social risk factors was zero (Table 7). Correlation coefficients between RSMR with and without adjustment for these factors were near 1 (0.9987 for dual-eligible, 0.9983 for low SES patients). Scatter plots showing this relationship are provided in Figures 6 and 7. This indicates that including these social risk factors in hospital-level measure scores result in limited differences in hospitals' measure results after accounting for other factors (age, comorbidities) included in the risk model.

Table 7: RSMR Distributions by Social Risk Factor

Risk Factor	Overall Hospital RSMR (with risk factor in the model)	Median Absolute Change in RSMR (%) (+/- risk factor)	Interquartile Range	Minimum	Maximum	Correlation Coefficient between RSMRs (+/- risk factor)
Dual-Eligible	0.0682	0.0000	0.0002	-0.0017	0.0033	0.9987
Low SES (AHRQ Index)	0.0682	0.0000	0.0001	-0.0089	0.0013	0.9983

Figure 6. Correlation of RSMR with and without Adjustment for Dual-Eligibility

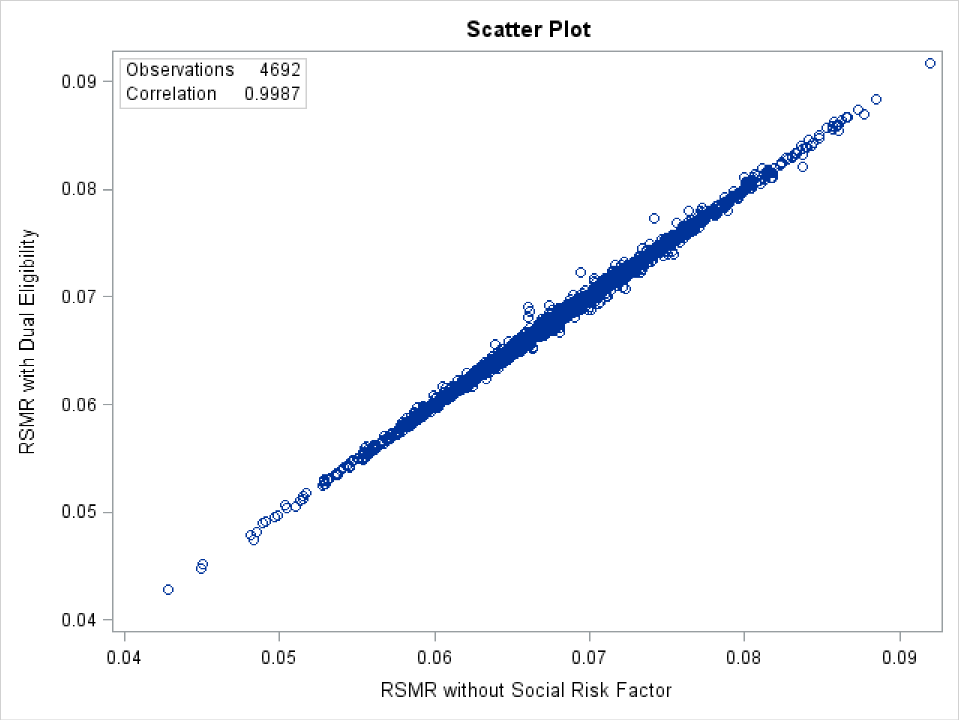
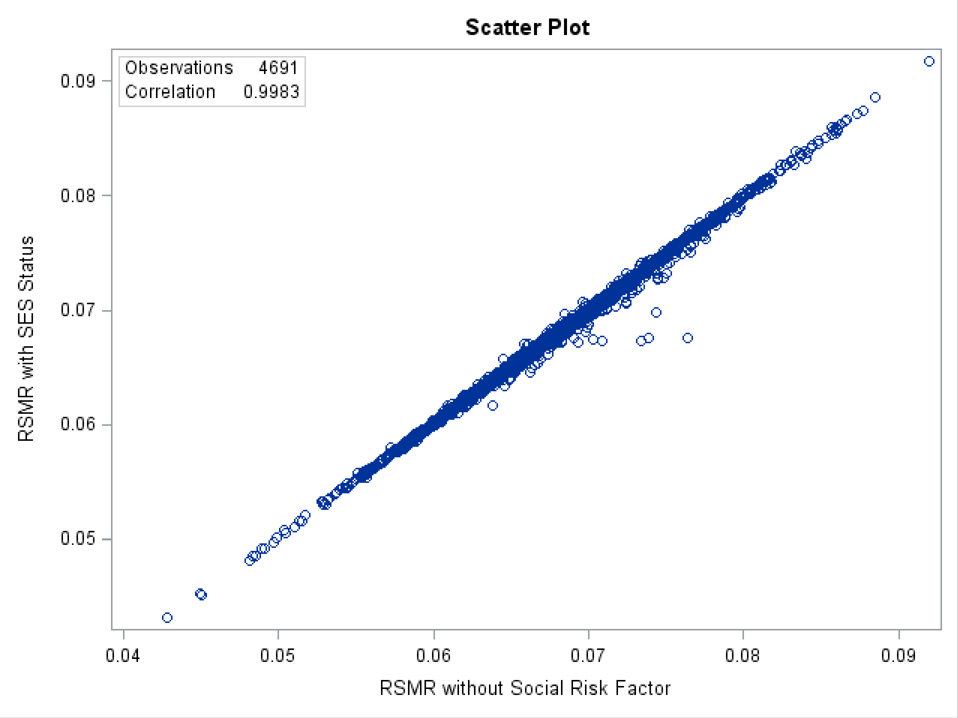


Figure 7. Correlation of RSMR with and without Adjustment for Low SES Status



In summary, we conclude that adjusting for social risk factors in this measure would have little effect on hospitals' measure scores:

• Correlation coefficients of measure scores comparing models with and without the social risk variables are near 1.0.
• C-statistics with the social risk variables in vs. out of the model, are unchanged.
• For most clinical divisions, neither the dual-eligible nor low AHRQ SES variables had a statistically significant association with the risk of mortality in a multivariate model.
• In the surgical cancer division, which did show a relationship with the outcome in a multivariate model for the AHRQ SES variable, we show that:

o There is a hospital-level effect and a patient-level effect of similar magnitude
o The hospital-level effect is larger compared with the hospital-level effect for the condition-based (comorbid) risk variables.
o This division represents just 2.0% of the total cohort.

Therefore, while adjusting for this variable would not have much impact, any adjustment would also remove hospital-level effects that may reflect lower-quality care provide to patients with low SES status.

Based on these results, the measure does not adjust for these social risk factors.

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

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

Discrimination Statistics

(1) Area under the receiver operating characteristic (ROC) curve (c-statistic)

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. To calculate the c-statistic, observed mortality rates were compared to predicted mortality probabilities across predicted rate deciles.

(2) Predictive ability

Discrimination in predictive ability measures the ability to distinguish high-risk subjects from low-risk subjects; therefore, for a model with good predictive ability we would expect to see a wide range in mortality rates between the lowest decile and highest decile. To calculate the predictive ability, we calculated the range of observed mortality rates between the lowest and highest predicted deciles.

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. Estimated calibration values of γ0 far from 0 and estimated values of γ1 far from 1 provide evidence of over-fitting.

We tested the performance of the model using Dataset 3, 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](#)

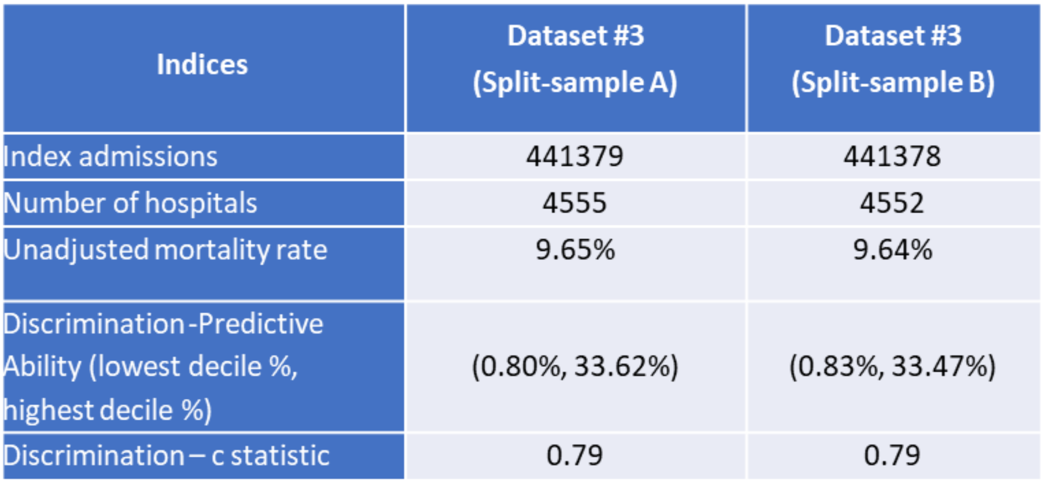
2b3.6. Statistical Risk Model Discrimination Statistics (*e.g., c-statistic, R-squared*)**:**

Table 8. Division-level C-Statistics

	Division	C-statistics (2016-2017)
Non-Surgical Divisions	Cancer	0.78
	Cardiac	0.84
	Gastrointestinal	0.84
	Infectious disease	0.84
	Neurology	0.83
	Orthopedics	0.82
	Pulmonary	0.81
	Renal	0.78
	Other Conditions	0.82
Surgical Divisions	Cancer	0.85
	Cardiothoracic	0.83
	General	0.87
	Neurosurgery	0.92
	Orthopedic	0.91
	Other Surgical Procedures	0.87

Table 9. Division-level Model Discrimination in the Non-surgical Pulmonary Division

Note: The Non-surgical Pulmonary Division is used here as an example. Predictive ability for all 15 models is available in the Data Dictionary tab “HWM_Pred_Abil”.



Indices	Dataset #3 (Split-sample A)	Dataset #3 (Split-sample B)
Index admissions	441379	441378
Number of hospitals	4555	4552
Unadjusted mortality rate	9.65%	9.64%
Discrimination-Predictive Ability (lowest decile %, highest decile %)	(0.80%, 33.62%)	(0.83%, 33.47%)
Discrimination – c statistic	0.79	0.79

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

Table 10. Statistical Risk Model Calibration Statistics

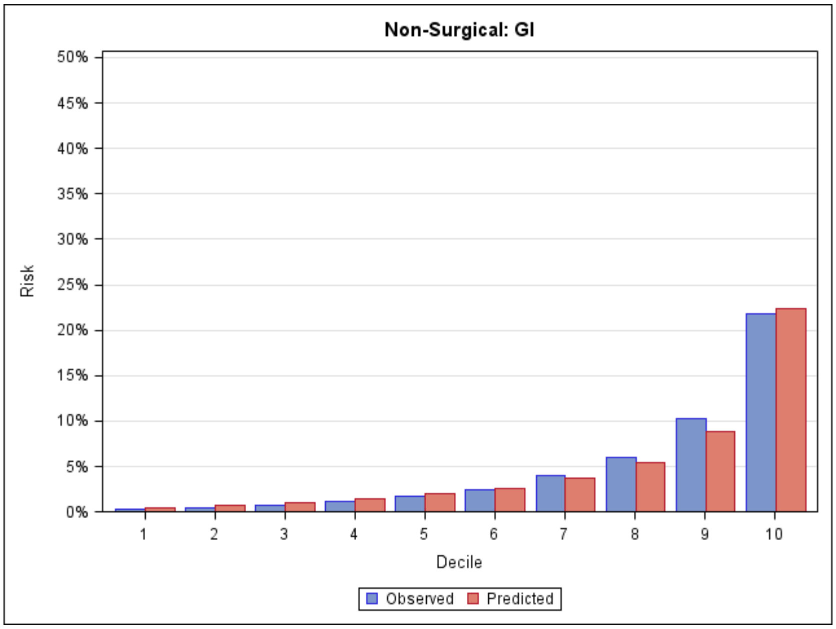
Division		Dataset #1 (Initial Development Dataset) (2014-2015)	Dataset #2 (ICD-10 re-specification dataset) (2016-2017)
Non-Surgical	Cancer	(0, 1)	(0.009, 0.995)
	Cardiac	(0, 1)	(-0.17, 0.946)
	Gastrointestinal	(0, 1)	(0.03, 1.014)
	Infectious Disease	(0, 1)	(-0.033, 0.982)
	Other Conditions	(0, 1)	(-0.019, 0.990)*
	Neurology	(0, 1)	(-0.052, 0.987)
	Orthopedics	(0, 1)	(-0.007, 0.992)
	Pulmonary	(0, 1)	(-0.023, 0.989)
	Renal	(0, 1)	(0.028, 1.01)
Surgical	Cancer	(0,1)	(-0.088, 0.976)
	Cardiothoracic	(0,1)	(-0.011, 0.995)
	General	(0,1)	(-0.021,0.987)*
	Other Surgical Procedures	(0,1)	(-0.055, 0.975)
	Neurosurgery	(0,1)	(0.01, 0.972)
	Orthopedics	(0,1)	(-0.051, 0.982)

*Two of the divisions had different diagnosis codes in the split samples and therefore some of the diagnosis codes did not have corresponding parameter estimates. To balance the CCS categories within each split sample we removed 1985 patients in total across both split samples.; for surgical general, we removed 283 patients. Prior to removing these patients (with unbalanced CCSs between the two split samples), the values of γ0 estimated values of γ1 were (-0.771, 0.731) for the non-surgical "other" division, and (-0.645, 0.733) for the surgical "general" division.

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

Please see the data dictionary, tab "HWM_Risk_Decile_Plots" for the risk decile plots for each of the 15 models. One representative example is shown below.

Figure 8. Risk Decile Plot for the Non-Surgical GI Division



2b3.9. Results of Risk Stratification Analysis:

Not applicable. This measure is not risk stratified.

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

The range of c statistic results is 0.75 to 0.91 which is consistent with or better than results we have seen for other 30-day mortality measures. The predictive ability results demonstrate a wide range between the lowest decile and highest decile for each of the 15 models, showing that that each model can distinguish between high and low-risk subjects. The risk-decile plots show that the predicted risk closely approximated the observed risk in most deciles, suggest good calibration.

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*)

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)***

We characterize the degree of variability by displaying and reporting the distribution of the RMSR.

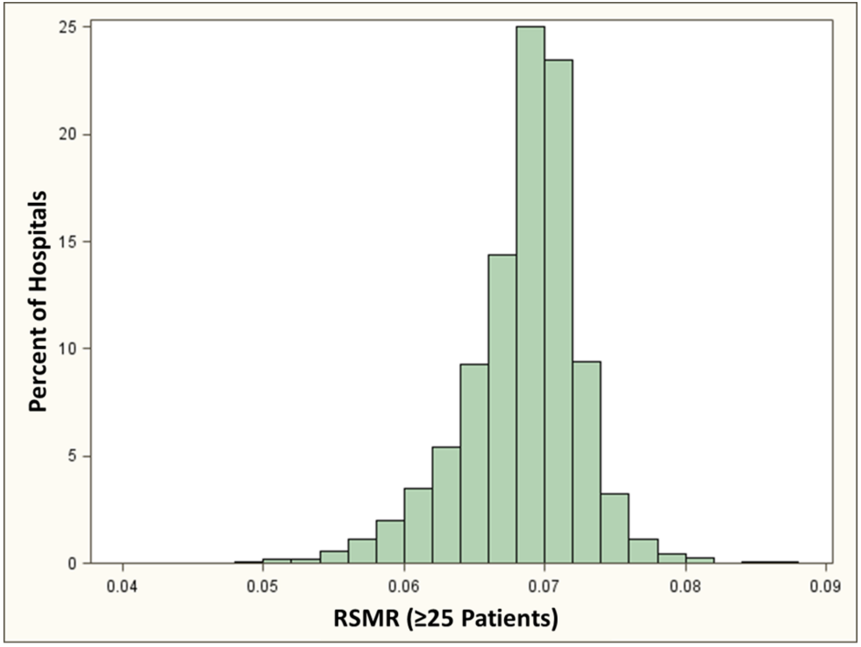
2b4.2. What were the statistical results from testing the ability to identify statistically significant and/or clinically/practically meaningful differences in performance measure scores across measured entities? **(e.g., *number and percentage of entities with scores that were statistically significantly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

Percentiles of distribution for the overall measure score (RSMR) for hospitals with at least 25 patients are shown in Table 11. The distribution of the measure score is shown in Figure 9.

Table 11. Distribution of RSMR

Percentile	Min	1st	5th	10th	25th	50th	75th	90th	95th	99th	Max
RSMR	3.95%	5.57%	6.07%	6.32%	6.66%	6.93%	7.09%	7.26%	7.40%	7.75%	8.70%

Figure 9. Distribution of the measure score (RSMR)



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 performance between the lowest-performing hospitals (RSMR of 3.95%) and the highest performing hospitals (RSMR of 8.7%) shows there is a clear quality gap.

In terms of performance compared to the median (6.93%), some hospitals can achieve substantially lower overall risk-standardized mortality rates than the average-performing hospital, while other hospitals are performing substantially worse than an average performer.

Specifically, the best performing hospital (RSMR of 3.95%) is performing 43% better than an average performer (or has about 30 fewer deaths per 1000 patients compared to the average performer), while the worst performing hospital (8.70%) is performing 25% worse than an average performer (or has 18 more deaths per 1000 patients). Note that the that average performer refers to hospital with the same case and service-line mix, performing at the average (median).

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

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

<u>**Note**</u>*: 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*)

Not applicable.

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*)**

Not applicable.

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*)

Not applicable.

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*)

Not applicable.

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*)

Not applicable.

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*)

Not applicable.

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 a combination of electronic sources

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

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.

N/A

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)
Public Reporting Not in use	

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

This measure is not yet in public reporting.

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?*)

See 4.a.1.3 below.

4a1.3. If not currently publicly reported OR used in at least one other accountability application, provide a credible plan for implementation within the expected timeframes -- any accountability application within 3 years and publicly reported within 6 years of initial endorsement. (*Credible plan includes the specific program, purpose, intended audience, and timeline for implementing the measure within the specified timeframes. A plan for accountability applications addresses mechanisms for data aggregation and reporting.*)

This measure may be included in the Hospital Inpatient Quality Reporting Program. CMS signaled, in the Inpatient Prospective Payment System (IPPS) proposed rule, the eventual possibility of including this measure (and/or the related hybrid measure) within the Inpatient Quality Reporting (IQR) program.

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

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

N/A; the measure is currently not in use.

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

N/A; the measure is currently not in use.

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

Describe how feedback was obtained.

N/A; the measure is currently not in use.

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

N/A; the measure is currently not in use.

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

N/A; the measure is currently not in use.

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

N/A; the measure is currently not in use.

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.

This is a new measure and there is no information available on performance improvement. This measure is not currently used in a program, but a primary goal of the measure is to provide information necessary to implement focused quality improvement efforts. Once the measure is implemented, we plan to examine trends in improvements by comparing RSMRs over time.

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 The measure is currently not in use.

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

N/A The measure is currently not in use.

5. Comparison to Related or Competing Measures

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

5. Relation to Other NQF-endorsed Measures

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

Yes

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

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

Hospital-Wide All-Cause Risk-Standardized Readmission Measure (NQF #1789);

Hospital-level risk-standardized complication rate (RSCR) following elective primary total hip arthroplasty (THA) and/or total knee arthroplasty (TKA) (NQF #1550);

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

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

Hospital 30-Day, all-cause, risk-standardized mortality rate (RSMR) following coronary artery bypass graft (CABG) Surgery (NQF #2558);

Hospital 30-day, all-cause, risk-standardized mortality rate (RSMR) following acute myocardial infarction (AMI) hospitalization (NQF #0230);

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

Hospital 30-day, all-cause, risk-standardized mortality rate (RSMR) following acute ischemic stroke hospitalization.

Death Rate in Low Mortality Diagnosis Related Groups (PSI-02) (NQF #0347)

AHRQ's Mortality for Select Conditions (IQI-90) (NQF #0530)

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.

This claims-only hospital-wide mortality (HWM) measure is intended to complement the existing CMS Hospital-Wide All-Cause Risk-Standardized Readmission Measure (NQF #1789) to allow assessment of trends in hospital performance for both readmission and mortality outcomes, similar to other complementary pairs of readmission and mortality measures for specific conditions and procedures. By measuring mortality outcomes across almost all hospitalized patients, this measure will provide an important additional performance assessment that will complement condition- and procedure-specific or other more narrowly defined mortality measures and allow a greater number of patients and hospitals to be evaluated. This HWM measure captures a similarly broad cohort to the CMS Hospital-Wide All-Cause Risk-Standardized Readmission Measure (NQF #1789), and a broader cohort than those of other CMS condition-specific measures. Because the mortality measure is focused on a different outcome, it differs from the existing CMS Hospital-Wide All-Cause Risk Standardized Readmission Measure (NQF #1789) in a couple of ways. First, this HWM measure includes patients with a principal discharge diagnosis of cancer (with some exceptions), whereas those patients are not included in the readmission measure. Cancer patients are appropriate to include in the HWM measure as many have survival as their primary goal; however due to cancer treatment plans, readmissions are frequently part of the plan and expected and therefore, are not a reasonable signal of quality. Another difference between the two measures is the number of divisions or specialty cohorts the patients are divided into, to more accurately risk adjust for case-mix and service-mix. The readmission measure divides patients into five categories, or "specialty cohorts", while the mortality measure uses 15. This is because the risk of mortality is much more closely related to patient factors than readmission is related to patient factors. PSI-02 (NQF #0357) is another complementary mortality measure, which captures a different patient population and a different outcome compared with the HWM measure submitted with this application. PSI-02 captures patients 18 years of age or older, or obstetric patients, whereas the HWM measure captures patients between the ages of 65 and 94. PSI-02 captures DRGs with less than 0.5% mortality rate, whereas the HWM measure captures all patients within all CCSs, regardless of mortality rate. Hospital-wide mortality captures mortality up to 30 days past admission, where AHRQ PSI-02 only captures in-hospital mortality. IQI 90 (NQF #0530) is another complimentary mortality measure, which is a composite measure of the number of in-hospital deaths for a narrow range of conditions (CHF, stroke, hip fracture, pneumonia, acute myocardial infarction and GI hemorrhage). The HWM measure presented in this application captures all deaths after 30 days of admission, for all conditions and procedures.

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.)

There are no competing NQF-endorsed measures.

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.

No appendix **Attachment:**

Contact Information

Co.1 Measure Steward (Intellectual Property Owner): Centers for Medicare & Medicaid Services (CMS)

Co.2 Point of Contact: Lein, Han, lein.han@cms.hhs.gov, 410-786-0205-

Co.3 Measure Developer if different from Measure Steward: Yale New Haven Health Services Corporation/Center for Outcomes Research and Evaluation (YNHHSC/CORE)

Co.4 Point of Contact: Karen, Dorsey, karen.dorsey@yale.edu, 203-764-5700-

Additional Information

Ad.1 Workgroup/Expert Panel involved in measure development

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

CORE convened a Technical Expert Panel comprised of clinicians, patients, and experts in quality improvement to provide input on key methodological decisions.

TEP Members:

- Michelle Beck – University of MD Upper Chesapeake Medical Center, MD
- Jeanne Black, PhD, MBA – Manager of Health Policy and Program Evaluation, Cedars-Sinai Health System; Los Angeles, CA
- John Bott – Manager of Healthcare Ratings, Consumer Reports; Yonkers, NY
- Roger Dmochowski, MD, MMHC, FACS – Executive Medical Director of Quality, Safety, and Risk, Vanderbilt University Medical Center; Nashville, TN
- Richard Dutton, MD, MBA – Chief Quality Officer, United States Anesthesia Partners; Houston, TX
- Gaye Hyre – Council Member and Patient/Family Caregiver Representative, CT State Innovation Model for Healthcare Equity and Access Council; Hartford, CT
- Irene Katzan, MD, MS – Director of Neurological Institute for Outcomes Research and Evaluation, Cleveland Clinic; Cleveland, OH
- Amy Kelley, MD, MSHA – Associate Professor and Staff Physician of Geriatrics and Palliative Medicine, Icahn School of Medicine at Mt Sinai; New York, NY
- Brenda Matti-Orozco, MD, FACP – Chief of Division of General Internal Medicine and Palliative Medicine and Hospice Medical Director, Morristown Medical Center and Atlantic Home Care & Hospice; Morristown, NJ
- Jyotirmay Sharma, MD, FACS – Associate Professor of Surgery and Medical Officer in Division of Healthcare Quality and Promotion, Emory University School of Medicine and Centers for Disease Control; Atlanta, GA

CORE convened two work groups comprised of patients and family caregivers, who represent important perspectives from diverse backgrounds, including a mix of genders, geographic locations, and experiences. They were essential in providing feedback around cohort development (particularly around whether to include hospice patients), and measure usability. We have withheld names to protect confidentiality.

Patient and Family Caregiver Work Group Members:

- Female, health advocate, and podcaster who has more than two decades of lived experience with complex chronic illness in New York.
- Male, heart transplant patient, intensive care unit advisory council for patient and families in Washington.
- Female, works with patients and families at a hospital in New Jersey.
- Female, educational psychologist, trained in a hospital, who has advocated for children with special needs in hospitals and schools in Texas.
- Male, is on a hospital advisory council for patients and families in New Jersey.

- Female, patient and family advisor for a medical group in New York.
- Female, is on the intensive care unit advisory council for patient and families in Washington.
- Female, current caregiver, and former patient and family advisor to a primary care medical group in New York.
- Female, council member for patients and families for a hospital in Washington.

CORE convened a Technical Work Group comprised of clinicians, surgeons, and statistical experts as well as experts in the development and challenges of a hospital-wide mortality measure.

Technical Work Group Members:

- Lee Fleisher, MD – Chair of Department of Anesthesiology and Critical Care, University of Pennsylvania Health System; Vice-Chair of the Consensus Standards Advisory Committee; Co-Chair of the Surgery Standing Committee of the National Quality Forum
- Leora Horwitz, MD, MHS – Associate Professor in the Departments of Population Health and Medicine at New York University School of Medicine; founding Director of the Center for Healthcare Innovation and Delivery Science; New York University Langone Medical Center, and of the Division of Healthcare Delivery Science, Department of Population Health, New York University School of Medicine
- David Shahian, MD - Professor of Surgery at Harvard Medical School; Vice President of the Massachusetts General Hospital (MGH) Center for Quality and Safety; and Associate Director of the MGH Codman Center for Clinical Effectiveness in Surgery; NQF Board and Executive Committee member; Chair of The Society of Thoracic Surgeons (STS) Council on Quality, Research, and Patient Safety and the STS Quality Measurement Task Force.

The CORE measure development team meets regularly and is comprised of experts in medicine, quality outcomes measurement, and measure development.

CORE Measure Reevaluation Team:

- Doris Peter, PhD – Reevaluation Lead, CORE
- Amy Salerno, MD – Development Lead, CORE
- Karen Dorsey, MD – Reevaluation Division Director, CORE
- Lisa Suter, MD – New Measure Division Director, CORE
- Darinka Djordjevic, PhD – Project Manager, CORE
- Lynette Lines, MS, PMP – Previous Project Manager, CORE
- Shuling Liu, PhD – Lead Analyst, CORE
- Yongfei Wang, MS – Analyst, CORE
- Nicole Cormier, MPH – Additional Team Member, CORE
- Erica Norton, BS – Task Coordinator, CORE
- Fior Rodriguez, BS – Research Assistant, CORE
- Keith Loh, BS – Research Assistant, CORE
- Alex Ferrante, BS – Research Assistant, CORE
- Amanda Audette, BS – Research Assistant, CORE
- Rajvi Shah, Research, MPH Associate, CORE
- Julianne Ani, MPH – Research Associate, CORE
- Zhenqiu Lin, PhD – Statistical Consultant, Director of Data Management and Analytics, CORE
- Li Li, PhD – Supporting Analyst, CORE

- Xin Xin, MS - Supporting Analyst, CORE
- Jeph Herrin, PhD – Statistical Consultant, CORE
- Nihar Desai, MD, MPH – Clinical Consultant
- Jaqueline Grady, MS – Statistical Consultant and Associate Director of Data Management and Analytics
- Susannah Bernheim, MD, MHS – Director of Quality Measurement, CORE
- Harlan Krumholz, MD, MS – Senior Advisor and Director, CORE

Measure Developer/Steward Updates and Ongoing Maintenance

Ad.2 Year the measure was first released:

Ad.3 Month and Year of most recent revision:

Ad.4 What is your frequency for review/update of this measure? Annually

Ad.5 When is the next scheduled review/update for this measure? 03, 2020

Ad.6 Copyright statement:

Ad.7 Disclaimers:

Ad.8 Additional Information/Comments: