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

Purple text represents the responses from measure developers.

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

Brief Measure Information

NQF #: 0229

Corresponding Measures:

De.2. Measure Title: Hospital 30-Day, All-Cause, Risk-Standardized Mortality Rate (RSMR) Following Heart Failure (HF) Hospitalization

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

De.3. Brief Description of Measure: The measure estimates a hospital-level 30-day, all-cause, risk-standardized mortality rate for patients discharged from the hospital with a principal diagnosis of HF. Mortality is defined as death for any cause within 30 days after the date of admission for the index admission. CMS annually reports the measure for patients who are 65 years or older and enrolled in fee-for-service (FFS) Medicare and hospitalized in non-federal hospitals or are patients hospitalized in Veterans Health Administration (VA) facilities.

1b.1. Developer Rationale: The goal of this measure is to improve patient outcomes by providing patients, physicians, hospitals, and policy makers with information about hospital-level, risk-standardized mortality rates following hospitalization for HF. 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.

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

S.4. Numerator Statement: The outcome for this measure is 30-day all-cause mortality. We define mortality as death from any cause within 30 days from the date of admission for patients 65 and older hospitalized with a principal diagnosis of HF.

S.6. Denominator Statement: This claims-based measure is used for a cohort of patients aged 65 years or older.

The cohort includes admissions for patients aged 65 years and older discharged from the hospital with a principal discharge diagnosis of HF and with a complete claims' history for the 12 months prior to admission. The measure is publicly reported by CMS for those patients 65 years and older who are Medicare FFS or VA beneficiaries admitted to non-federal or VA hospitals, respectively.

Additional details are provided in S.7 Denominator Details.

S.8. Denominator Exclusions: The mortality measures exclude index admissions for patients:

1. Discharged alive on the day of admission or the following day who were not transferred to another acute care facility.
2. With inconsistent or unknown vital status or other unreliable demographic (age and gender) data.
3. Enrolled in the Medicare hospice program or used VA hospice services any time in the 12 months prior to the index admission, including the first day of the index admission.
4. Discharged against medical advice (AMA); or
5. Patients undergoing left ventricular assist device (LVAD) implantation or heart transplantation during an index admission or who have a history of LVAD or heart transplant in the preceding year.

For patients with more than one admission for a given condition each year, only one index admission for that condition is randomly selected for inclusion in the cohort for each year.

De.1. Measure Type: Outcome

S.17. Data Source: Claims, Enrollment Data, Other

S.20. Level of Analysis: Facility

IF Endorsement Maintenance – Original Endorsement Date: May 09, 2007 **Most Recent Endorsement Date:** Feb 19, 2016

IF this measure is included in a composite, NQF Composite#/title:

IF this measure is paired/grouped, NQF#/title:

De.4. IF PAIRED/GROUPED, what is the reason this measure must be reported with other measures to appropriately interpret results? This measure is paired with a measure of hospital-level, all-cause, 30-day, risk-standardized readmission (RSRR) following HF hospitalization.

Preliminary Analysis: Maintenance of Endorsement

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

Criteria 1: Importance to Measure and Report

1a. Evidence

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

1a. Evidence. The evidence requirements for a health outcome measure include providing empirical data that demonstrate a relationship between the outcome and at least one healthcare structure, process, intervention, or service.

Prior review in 2015

- This measure calculates hospitals' 30-day risk-standardized mortality rate for patients who have been hospitalized with heart failure (HF).
- As a rationale for measuring this health outcome, the developers suggest that hospitals can influence mortality rates through a broad range of clinical activities, including prevention of complications, provision of evidence-based care, discharge planning, management of care transitions, medication reconciliation, and patient education.
- The developer provided numerous studies demonstrating that: (1) Appropriate and timely treatment for HF patients can reduce the risk of mortality within 30 days of hospital admission; (2) Trials of interventions which improve patient education upon discharge have been shown to improve survival for HF patients; and (3) Hospitals have been able to reduce mortality rates through these quality-of-care initiatives illustrates the degree to which hospital practices can affect mortality rates.
- The developer states that 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.

Changes to evidence from last review

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

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

Updates:

- The developer provided information on the lifetime risk, prevalence, and cost of HF.
- The developer provided new evidence tying coordinated care for HF patients to reductions in all-cause mortality after HF admission. Additional evidence provided strengthens support for the previously submitted conclusions.

Question for the Committee:

- The evidence provided by the developer is updated, directionally the same, and stronger compared to that for the previous NQF review. Does the Committee agree there is no need for repeat discussion and vote on Evidence?

Guidance from the Evidence Algorithm

The measure assesses performance on a health outcome of 30-day all-cause mortality (box 1) → The relationship between decreased risk of 30-day all-cause mortality and at least one process is demonstrated through empirical data (box 2) → Pass

Preliminary rating for evidence: ☒ Pass ☐ No Pass

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

Maintenance measures – increased emphasis on gap and variation

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

- The developer provided three-year, hospital-level, risk-standardized mortality rates (RSMRs) using Medicare claims and VA administrative data (1,081,897 admissions from 4,637 hospitals) from July 1, 2016 to June 30, 2019.
    - The RSMRs have a mean of 11.4%, a standard deviation of 1.6 and a range from 5.3 – 18.5%. The median risk-standardized rate is also 11.4%.

Disparities

- The developer included the distribution of 30-day HF RSMRs by Proportion of Dual Eligible Patients from July 2016 through June 2019.
- Description of Social Risk Variable//Dual Eligibility
    - Quartile // Q1 // Q4
- Social Risk Proportion (%) // (0-8.37) // (34.43-100)
    - # of Hospitals // 910 // 910
    - 100% Max // 17.7 // 18.5
    - 90% // 13.6 // 13.4
    - 75% // 12.7 // 12.3
    - 50% // 11.5 // 11.2
    - 25% // 10.3 // 10.1
    - 10% // 9.4 // 9.2
    - 0% Min // 6.7 // 5.3
- The developer also included the distribution of 30-day HF RSMRs by Proportion of Patients with AHRQ SES Index Scores from July 2016 through June 2019.
- Description of Social Risk Variable //AHRQ SES Index
    - Quartile // Q1 // Q4
- Social Risk Proportion (%) // (0-10.24)// (23.59-100)
    - # of Hospitals // 911 // 911
    - 100% Max // 17.1 // 17.1
    - 90% // 13.3 // 13.4
    - 75% // 12.3 // 12.3
    - 50% // 11.1 // 11.1
    - 25% // 10.0 // 10.0
    - 10% // 9.0 // 9.0
    - 0% Min // 6.1 // 5.3

Questions for the Committee:

- Is there a gap in care that warrants a national performance measure?
- Are you aware of evidence that additional disparities exist in this area of healthcare aside from what the developer provided?

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

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability: Specifications and Testing

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

Reliability

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

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

Validity

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

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

Composite measures only:

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

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

Evaluators: NQF Scientific Methods Panel Subgroup

Methods Panel Review (Combined)

Methods Panel Evaluation Summary:

This measure was reviewed by the Scientific Methods Panel (SMP). The SMP Subgroup passed the measure on reliability and validity. The measure was not pulled for discussion during the October 2020 meeting. A summary of the measure and the SMP's review is provided below.

Reliability

- Measure passed the SMP review for reliability with moderate rating (H-4; M-4; L-3; I-0)
- The developers conducted two types of reliability testing. The developers estimated measure score level by calculating the intra-class correlation coefficient (ICC) using a split sample method (i.e., test-retest), and then estimated the facility-level reliability (signal-to-noise reliability)
    - Using the Spearman-Brown prediction formula, the developers estimated that the agreement between the two independent assessments of the RSMR for each hospital with 25 admissions was 0.632.

- The median reliability (signal-to-noise) score was 0.79, ranging from 0.34 to 0.99, and the 25th and 75th percentiles were 0.58 and 0.9, respectively, for the signal-to-noise testing for each hospital with at least 25 admissions

- Most SMP members agreed that the reliability tests were appropriate and that the results show moderate reliability. One member voiced concerns about low reliability for the bottom 10% hospitals in the signal-to-noise ratio analysis (r<0.44) and split-sample reliability (0.63), stating this was acceptable but not ideal.
- In response to the concerns and questions raised, the developer clarified that the 25-case minimum is established by CMS and is aligned across all mortality and readmission measures for public reporting.

Validity

- Measure passed the SMP review for validity with moderate rating (H-0; M-6; L-1; I-1)
- The developers conducted validity testing at the performance measure score level, including both empirical validity testing (by comparing CMS' Star-Rating Mortality Scores and Star Rating summary scores), and systematic assessment of face validity
    - The correlation between HF RSMRs and the Star-Rating mortality score was -0.676, which suggests that hospitals with lower HF RSMRs are more likely to have higher Star-Rating mortality scores
    - The correlation between HF RSMRs and the Star-Rating summary score was -0.114, which suggests that hospitals with lower HF RSMRs are more likely to have higher Star-Rating summary scores
    - The risk model includes 24 clinical and demographic risk factors. Dual eligibility and AHRQ SES index were tested but not included in the final model.
    - The developers noted that the addition of any of these variables into the hierarchical model has little to no effect on hospital performance (c-statistic remains 0.73). The developer showed that there was little impact on measure scores as gauged by the difference between measure scores calculated with versus without the social risk factors in the model.
- The members voiced no concerns about validity and noted that the exclusions are appropriate.

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**

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

2a1. Reliability-Specifications: Which data elements, if any, are not clearly defined? Which codes with descriptors, if any, are not provided? Which steps, if any, in the logic or calculation algorithm or other specifications (e.g., risk/case-mix adjustment, survey/sampling instructions) are not clear? What concerns do you have about the likelihood that this measure can be consistently implemented?

- This is a complex measure that was reviewed by the Scientific Methods Committee. It is a measure that has been used since 2007. It uses Medicare claims data that is in discrete data fields. I think it can be implemented consistently, however the SMC did have some concerns about the risk adjustment model - 1) model was based on data from 1998 and may not reflect contemporary care, 2) the lack of inclusion of social risk factors in the model.
- Passed SMP review for reliability and validity
- No questions
- I thought that most of this was clearly stated and defined. I do not recall seeing adequate definitions of race/ethnicity, insurance, or disability (nor outcomes by these subpopulations).
- Agree with/Defer to SMP
- The specifications are clear, and the measure developer does a good job describing hierarchical logistic regression models. Some additional details on how to is used for risk adjustment may be beneficial.
- This measure was reviewed by the Scientific Methods Panel. The Subgroup passed the measure on reliability. I do not see a reason to challenge their conclusion.
- no concerns
- no concerns

2a2. Reliability - Testing: Do you have any concerns about the reliability of the measure?

- The Split Sample reliability correlation was acceptable at 0.632; I was a little concerned about the range of Signal-to-Noise results from 0.34 to 0.99 (median 0.79)
- No
- No concerns
- No
- Agree with/Defer to SMP
- No concerns - reviewed by NQF Scientific Methods Panel
- No
- no concerns
- no concerns

2b1. Validity -Testing: Do you have any concerns with the testing results?

- Clinical validity - a chart-based model was compared to an administrative claims-based model. The administrative claims model had an accuracy of 69-71% and the chart-derived model had an accuracy of 75-78%. Both seem a little low but also the data set was from 1998. This was a concern for the Scientific Method Committee. Empiric validity was tested by comparing RSMR to the Hospital-Compare Star-Rating overall mortality. Since the HF mortality would be a part of overall mortality for the institution there should be a correlation. There were also some concerns about the risk adjustment model: 1) calibration data from 1998 and 2) not including the social risk factors in the model. (See p. 15)
- No
- No concerns

- No
- Agree with/Defer to SMP
- No concerns - reviewed by NQF Scientific Methods Panel
- This measure was reviewed by the Scientific Methods Panel. The Subgroup passed the measure on validity. I do not have concerns.
- It was surprising that the correlation between HF RSMRs and Star-Rating summary score was low at -0.114 despite moderate correlation with Star-Rating mortality score at -0.676.
- no concerns

2b2-3. Other Threats to Validity (Exclusions, Risk Adjustment) 2b2. Exclusions: Are the exclusions consistent with the evidence? Are any patients or patient groups inappropriately excluded from the measure? 2b3. Risk Adjustment: If outcome (intermediate, health, or PRO-based) or resource use performance measure: Is there a conceptual relationship between potential social risk factor variables and the measure focus? How well do social risk factor variables that were available and analyzed align with the conceptual description provided? Are all the risk-adjustment variables present at the start of care (if not, do you agree with the rationale provided)? Was the risk adjustment (case-mix adjustment) appropriately developed and tested? Do analyses indicate acceptable results? Is an appropriate risk-adjustment strategy included in the measure?

- See answer 7 2b1. Scientific Methods Committee - more than one member questioned why the social risk factors were not included in the model when the odds ratios for them were similar to other factors included in the model.
- Exclusions appropriate
- Risk adjustment appears appropriate
- I am not concerned about exclusions. I do not think Table 4 is adequately explained. Where is race/ethnicity?
- None
- Additional data on how risk adjustment is applied may be beneficial.
- The discrepancy between the SDS outcomes based on the variables that are available (minor impact of SDS) and what clinicians believe they are observing in their practices is a perpetual and unsolved problem. I think that the clinicians are observing the impact of social isolation, a factor not captured in the medical record.
- Risk adjustment strategy was included in the measure with acceptable results, c-statistic 0.73. Addition of SES variables did not significantly change prediction of the model.
- no concerns

2b4-7. Threats to Validity (Statistically Significant Differences, Multiple Data Sources, Missing Data) 2b4. Meaningful Differences: How do analyses indicate this measure identifies meaningful differences about quality? 2b5. Comparability of performance scores: If multiple sets of specifications: Do analyses indicate they produce comparable results? 2b6. Missing data/no response: Does missing data constitute a threat to the validity of this measure?

- No concerns
- N/A No missing data
- No questions
- I find nothing of concern beyond the Scientific Methods Panel.
- My only question is whether this measure shows performance gaps or is just a proxy for the population served.
- No concerns

- I do not see significant threats to validity
- no significant threats to validity identified, split sample analyses produced reliable results
- 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 the data elements originate from defined fields in electronic claims.
    - The necessary data are coded by someone other than the person obtaining the original information (e.g., DRG, ICD-9 codes on claims).
    - This measure uses administrative claims data and enrollment data and as such, it offers no data collection burden to hospitals or providers.

Questions for the Committee:

- 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

Feasibility: Which of the required data elements are not routinely generated and used during care delivery? Which of the required data elements are not available in electronic form (e.g., EHR (Electronic Health Record) or other electronic sources)? What are your concerns about how the data collection strategy can be put into operational use?

- All data reported to be available in electronic claims. Existing measure used for many years
- This measure, as indicated, uses administrative claims data and enrollment data. There is no burden to hospital or provider.
- Measure is in use
- What is the feasibility that all mortality data is captured? Might high social risk populations have less reliable mortality data? How accurately are local reports of death synchronized with the databased used. EHR data is often lacking in this.
- There is a lag in reporting be definition of the measure but appears feasible.
- No concerns with feasibility. This measure has been in use in a variety of measurement systems
- I agree that the measure is at least moderately feasible
- No concerns. Data can be extracted from EHR automatically with no additional human resource input
- no concerns

Criterion 4: Usability and Use

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

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

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

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

Current uses of the measure

Publicly reported? ☒ **Yes** ☐ **No**

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

Accountability program details

- The developer noted that the measure is publicly reported on CMS' Care Compare website. Under Care Compare, CMS collects quality data from hospitals, with the goal of driving quality improvement through measurement and transparency by publicly displaying data to help consumers make more informed decisions about their health care. It is also intended to encourage hospitals and clinicians to improve the quality and cost of inpatient care provided to all patients. The data collected are available to consumers and providers on the Hospital Compare website.
- The developer noted that the measure is also within CMS' Hospital Value-Based Purchasing (HVBP) 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

1. Those being measured receive performance results and data via CMS's QualityNet website. The website also contains detailed patient-level results and benchmarks to assist in interpretation.
2. The developer noted that measured entities can submit feedback about the measure through an email inbox. Experts on measure specifications, calculation, or implementation, prepare responses to those inquiries and reply directly to the sender.
3. The developers state that they consider feedback when reevaluating measures. The developers state that they have not received any feedback from stakeholders that would require additional analysis or changes to the measure since the last endorsement maintenance cycle.

Questions for the Committee:

How has the measure been vetted in real-world settings by those being measured or others?

Preliminary rating for Use: ☒ Pass ☐ No Pass

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

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

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

Improvement results

- The developers reported the median hospital 30-day, all-cause, RSMR for the HF mortality measure for the 3-year period between July 1, 2016 and June 30, 2019 was 11.4%. The median RSMR decreased by 0.7 absolute percentage points from July 2016-June 2017 (median RSMR: 11.6%) to July 2018-June 2019 (median: RSMR: 10.9%).

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

Unexpected findings (positive or negative) during implementation

- The developers report they have not seen any unexpected findings.

Potential harms

- The developer noted that providers could inappropriately shift care in response to this measure. They monitor for this unintended consequence and have not seen any indications it is occurring.

Additional Feedback: N/A

Questions for the Committee:

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

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

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

4a1. Use - Accountability and Transparency: How is the measure being publicly reported? Are the performance results disclosed and available outside of the organizations or practices whose performance is measured? For maintenance measures - which accountability applications are the measure being used for? For new measures - if not in use at the time of initial endorsement, is a credible plan for implementation provided? 4a2. Use - Feedback on the measure: Have those being measured been given performance results or data, as well as assistance with interpreting the measure results and data? Have those being measured or other users been given an opportunity to provide feedback on the measure performance or implementation? Has this feedback has been considered when changes are incorporated into the measure?

- Publicly reported in Hospital Compare. Developer reports that there is an established mechanism for feedback.
- Yes, no concerns
- Feedback on performance is available publicly
- These data are widely available and publicized.
- Does CMS Hospital compare also list social determinants of health to help consumers put this in context?
- Measure is in use. Data is meaningful
- The measure is publicly reported on Hospital Compare. Nobody seems to complain about it.

- Measure is available publicly through hospital compare website. Feedback on the measure provided via email.
- no concerns

4b1. Usability – Improvement: How can the performance results be used to further the goal of high-quality, efficient healthcare? If not in use for performance improvement at the time of initial endorsement, is a credible rationale provided that describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations? 4b2. Usability – Benefits vs. harms: Describe any actual unintended consequences and note how you think the benefits of the measure outweigh them.

- Improvement: Median RSMR has only decreased 0.7% comparing the 1st year analyzed to the 3rd year analyzed.
- Continued dissemination of results. Benefit of measure outweighs unintended consequences. The potential for hospitals not to admit HF patients is not something I am seeing, but rather with the measure implementation of best practices to prevent 30-day HF readmission and mortality.
- Benefits likely outweigh harms
- I was also concerned about increased mortality from hospitals avoiding readmissions, but studies have not studies have not associated lower readmission rates with increased mortality.
- No harms evident, unclear whether this measure is actionable.
- Data is usable. Of note, due to COVID-19, is will be interesting how this measure looks year-to-year
- There has been some improvement in RSMR between July 2016-17 and 2018-2019. No harms have been identified.
- Data was used to encourage transparency in hospital performance. It is also used in the CMS's Hospital Value-Based Purchasing (HVBP) Program. Significant improvements on the measure noted over time. No significant unintended harm reported.
- No concerns

Criterion 5: Related and Competing Measures

Related or competing measures

- 0330 Hospital 30-Day, All-Cause, Risk-Standardized Readmission Rate (RSRR) Following Heart Failure (HF) Hospitalization
- 0358 Heart Failure Mortality Rate (IQI (Inpatient Quality Indicator) 16)
- 0468 Hospital 30-Day, All-Cause, Risk-Standardized Mortality Rate (RSMR) Following Pneumonia Hospitalization
- 1789 Hospital-Wide All-Cause Unplanned Readmission Measure (HWR (Hospital Wide Readmission))
- 1893 Hospital 30-Day, All-Cause, Risk-Standardized Mortality Rate (RSMR) Following Chronic Obstructive Pulmonary Disease (COPD) Hospitalization
- 3502 Hybrid Hospital-Wide (All-Condition, All-Procedure) Risk-Standardized Mortality Measure
- 3504 Claims-Only Hospital-Wide (All-Condition, All-Procedure) Risk-Standardized Mortality Measure

Harmonization

- The developer stated that the measure specifications are harmonized to the extent possible. They noted that they focused on related outcome measures (mortality and readmissions) in their harmonization analysis. Their rationale for this was that clinical coherence of the measured cohort takes precedence over alignment with related non-outcome measures. They state that many process measures are limited due to the broader patient exclusions necessary to examine only a specific subset of patients who are eligible for that measure (e.g., patients who receive a specific medication or undergo a specific procedure).

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

Related and Competing: Are there any related and competing measures? If so, are any specifications that are not harmonized? Are there any additional steps needed for the measures to be harmonized?

- There are 7 related and competing measures, with two of them related to readmission rates. The others are for other diagnoses related to the pulmonary system or to mortality rates (hybrid hospital-wide and claim only hospital-wide)
- As listed, agree with the developer's statement on harmonization
- Measures have been harmonized
- They are related in that patients may have multiple active conditions in the same admission, notably MI (Myocardial Infarction) and CHF (Congestive Heart Failure), but I would not say they are 'competing'
- No issues.
- No concerns
- The developer stated that the measure specifications are harmonized to the extent possible. I do not see any evidence to dispute this assertion.
- Yes, several. Hospital 30-day, All-Cause, Risk-Standardized Mortality Rate (RSMR) Following Acute Myocardial Infarction (AMI) Hospitalization since some patients can be admitted with both diagnoses. The developers state that the measures were harmonized to extent possible.
- No concerns

Public and Member Comments

Comments and Member Support/Non-Support Submitted as of: 01/21/2021

- Comment by American Medical Association

  The American Medical Association (AMA) appreciates the opportunity to comment on #229, Hospital 30-day, all-cause, risk-standardized mortality rate (RSMR) following heart failure (HF) hospitalization. We are disappointed to see the minimum measure score reliability results of 0.34 using a minimum case number of 25 patients. We believe that measures must meet minimum acceptable thresholds of 0.7 for reliability.

  In addition, the AMA is extremely concerned to see that the measure developer used the recommendation to not include social risk factors in the risk adjustment models for measures that are publicly reported as outlined in the recent report to Congress by Assistant Secretary for Planning and Evaluation (ASPE) on Social Risk Factors and Performance in Medicare's Value-based Purchasing program (ASPE, 2020). We believe that while the current testing may not have produced results that would indicate incorporation of the two social risk factors included in testing, this measure is currently used both for public reporting and value-based purchasing. A primary limitation of the ASPE report was that none of the recommendations adequately addressed whether it was or was not appropriate to adjust for social risk factors in the same measure used for more than one accountability purpose, which is the case for here. This discrepancy along with the fact that the additional analysis using the American Community Survey is not yet released must be addressed prior to any measure developer relying on the recommendations within this report.

  We request that the Standing Committee evaluate whether the measure meets the scientific acceptability criteria.

Reference:

Office of the Assistant Secretary for Planning and Evaluation, U.S. Department of Health & Human Services. Second Report to Congress on Social Risk Factors and Performance in Medicare's Value-Based Purchasing Program. 2020. https://aspe.hhs.gov/social-risk-factors-and-medicares-value-based-purchasing-programs

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

- 0 support the measure
- 1 does not support the measure

Combined Methods Panel Scientific Acceptability Evaluation

Scientific Acceptability: Preliminary Analysis Form

Measure Number: 0229

Measure Title: Hospital 30-day, All-Cause, Risk-Standardized Mortality Rate (RSMR) following heart failure (HF) hospitalization

Type of measure:

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

☒ **Outcome** ☒ **Outcome: PRO-PM (Patient Reported Outcomes Performance Measures)** ☐ **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**

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.)

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.**

   Panel Member #1: No concerns

   Panel Member #2: None

Panel Member #4: The documentation provided in the MIF (Measure Information Form) file describing the numerator (S.5) and denominator (S.6) was a bit unclear and disorganized, especially in trying to reflect all the different possible populations (FFS 65+, VA, all-payer).

Panel Member #5: Overall, measure specifications are very clear.

There is one general concern. The specifications include patients aged 18+. However, it seems that most testing was conducted using data from patients aged 65+. The only testing conducted for patients aged 18-64 vs. 65+ was for the risk-adjustment model (section 2b3.11) using data from 2006. I could not identify any other testing of reliability, validity, threats to validity, or performance that included data for the younger age group. This questions the reliability, and possibly also the validity of this measure for patients aged 18-64. This issue has been clarified, and measure developers decided to change the specifications to limit each of the measures to the Medicare FFS 65+ population. The ratings for reliability and validity were selected accordingly.

A minor comment: On S.14, the sentence "The estimated hospital-specific intercept is added coefficients multiplied by the patient characteristics." seems to be missing "… is added to the sum of the estimated regression coefficients…."

Panel Member #6: Numerator is ambiguous: We define mortality as death from any cause within 30 days of the index admission date for patients 18 and older discharged from the hospital with a principal diagnosis of HF.

Does this include both patients who were discharged alive and patients who were discharged dead? If so, wording should be clarified, e.g., death from any cause within 30 days of the index admission date for patients 18 and older who have a principal diagnosis at discharge of HF.

Panel Member #7: No significant concerns. **"**This claims-based measure can be used in either of two patient cohorts: (1) patients aged 65 years or older, or (2) patients aged 18 years or older." "This measure is paired with a measure of hospital-level, all-cause, 30-day, risk-standardized readmission (RSRR) following HF hospitalization."

Panel Member #8: None

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**

   Panel Member #6: Although appropriate score level testing was performed, developers also note that effort was made to choose data element for risk modeling that are subject to CMS audit and have been shown in the past to have good reliability (no data)

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**

   Panel Member #1: N/A

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

   Submission document: Testing attachment, section 2a2.2

   Panel Member #1: Developer used a split sample ICC and signal to noise approaches, which were appropriate.

Panel Member #3: SNR (Signal to Noise Ratio) in validation set was 0.79. Split sample reliability score 0.63. Both are consistent with acceptable reliability

Panel Member #4: Used two appropriate methods for testing – split sample and signal-to-noise

Panel Member #5: No concerns. Methods were appropriate and clearly described. A description of how the 25-case threshold for public reporting was determined would be useful.

Panel Member #6: Split sample and signal-to-noise, appropriate

Panel Member #7: "We performed types of testing. First, we estimated the overall measure by calculating the intra-class correlation coefficient (ICC) using a split sample (i.e., test-retest) method. Second, we estimated the facility-level reliability (signal-to-noise reliability)."

Panel Member #8: Split sample ICC and signal to noise ratio

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

Submission document: Testing attachment, section 2a2.3

Panel Member #1: The STN analysis found a median facility (>25 admission) reliability estimate of 0.79. The split sample ICC demonstrated a reliability of 0.63. Both results indicate acceptable reliability.

Panel Member #2: Signal to noise wide range of reliability scores: 0.34 to 0.99 – Q1 values is 0.58 – moderate by most scales. Split sample ICC – 0.79 – substantial agreement

Panel Member #3: SNR in validation set was 0.79. Split sample reliability score 0.63. Both are consistent with acceptable reliability.

Panel Member #4: Median signal-to-noise score of 0.79, which is substantial agreement, as defined by Adams et al. Split-sample score was 0.632 was perhaps the lower end of substantial.

Panel Member #5: I do not think the interpretation of SNR reliability estimate as an agreement statistic is appropriate. Results suggest acceptable reliability at the score level (>0.7, median = 0.79), thus there is high/acceptable certainty that the performance measure scores are reliable. However, it is now known how the inclusion of patients below the age of 65 would have impacted these results.

It would be useful to report here the percent of hospitals included in the reliability results (25+ cases), although this is reported in the performance section (21%).

Panel Member #6: Split sample 0.668; signal to noise from 0.3198, average 0.73 – good range

Panel Member #7: Using the Spearman-Brown prediction formula, the agreement between the two independent assessments of the RSMR for each hospital was 0.632.

SNR,

25th Percentile	Median	75th Percentile
0.58	0.79	0.90

Panel Member #8: The signal to noise ratio analysis revealed that median reliability for hospitals with >25 admissions was 0.79 and the 25th percentile was 0.58. The reliability is quite good for most entities but concerningly low for the bottom 10% (r<0.44). The split sample reliability analysis revealed that the overall reliability was 0.63. Although the Landis modifiers are cited, I do not accept them as relevant to this context. The Landis modifiers pertain to the strength of evidence against the null hypothesis of no agreement between raters of a categorical classifier. Split sample reliability of 0.63 perhaps acceptable but certainly not ideal. Note that other modifiers exist Koo 2016 - "values less than 0.5, between 0.5 and 0.75, between 0.75 and 0.9, and greater than 0.90 are indicative of poor, moderate, good, and excellent reliability, respectively. Portney and Watkins are more conservative, particularly at the upper end, with <0.75 poor to moderate, >0.75 good, an >0.90 ''reasonable for clinical measurements''.

I think we really need to move beyond these modifiers and do some work on the implications of unreliability in different quality measurement contexts. Can the developers comment of the impact of the observed reliability on misclassification or other consequences?

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** (considering 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.**

    Panel Member #1: The results indicate acceptable reliability, although the ICC reliability estimate of 0.63 is modest.

    Panel Member #2: See #7

    Panel Member #3: SNR in validation set was 0.79. Split sample reliability score 0.63. Both are consistent with acceptable reliability

    Panel Member #4: Used two appropriate methods for testing; signal-to-noise produced a score that would be categorized as 'substantial' agreement.

    Panel Member #5: Results suggest acceptable reliability at the score level, thus there is high certainty that the performance measure scores are reliable. Can developers elaborate on how the 25-case threshold was established in relation to the overall reliability results?

    Panel Member #6: Appropriate testing with good results; ambiguity of wording of definition of concern but easily addressed

    Panel Member #7: By the numbers.

    Panel Member #8: See my comments under #7

VALIDITY: ASSESSMENT OF THREATS TO VALIDITY

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

    Submission document: Testing attachment, section 2b2.

Panel Member #1: None.

Panel Member #2: None.

Panel Member #3: none

Panel Member #4: None

Panel Member #5: No concerns. Most cases excluded were due to being discharged alive on the day of admission or the following day who were not transferred to another acute care facility (4.2%) which is a criterion that has strong face validity and does not require additional testing. Other exclusions were less frequent (<1.4%) and have strong face validity.

Panel Member #6: As noted, ambiguity of the definition raises the question as to whether in-hospital deaths are included or excluded—this needs to be clarified.

Panel Member #7: None

Panel Member #8: None

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

    Submission document: Testing attachment, section 2b4.

    Panel Member #1: None.

    Panel Member #3: none

    Panel Member #4: There is variation in the calculated RSMRs; the statistical choice of how to categorize hospitals into 3 performance categories leaves 90% of hospitals in "no different from the U.S. national rate", which reflects some variation.

    Panel Member #5: As noted above, a clarification about the patient level performance transformation would be helpful: "The results are then transformed and…".

    As reported, 21% (992/4637) of hospitals had fewer than 25 cases therefor could not be reliably assessed for their RSMR (risk-standardized mortality rate). Can developers elaborate on how the 25-case threshold was established?

    Panel Member #6: Nice distribution of outcome measure across hospitals

    Panel Member #7: None

    Panel Member #8: None

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.

    Panel Member #1: N/A

    Panel Member #4: Not applicable.

    Panel Member #6: Multiple data sources used to confirm member eligibility and death. Methodology for linking of databases is well-established.

    Panel Member #7: None

    Panel Member #8: None

15. **Please describe any concerns you have regarding missing data.**

    Submission document: Testing attachment, section 2b6.

    Panel Member #1: None

    Panel Member #4: No missing data

    Panel Member #5: No concerns – no missing data reported

Panel Member #6: Exclusions for missing data do not seem to seriously bias results.

Panel Member #7: None

Panel Member #8: None

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

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

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**

Panel Member #1: The risk adjustment approach is sound, and results are acceptable, however I have a concern about the age of the data used to derive the coefficients in the model.

Panel Member #2: Methodology and results acceptable

Panel Member #3: Model discrimination (C stat in test is 0.69). Model calibration in validation data set is acceptable: (-0.004, 0.99). Feedback developers: calibration statistics are based on data from 1998, which is nearly 20 years old. MD needs to examine calibration in the testing data sets (2016-2019) using the models that have been updated for this time period. Risk adjustment model is unlikely to reflect contemporary care because it was developed using data from 1998. This is a major threat to measure validity.

Panel Member #4: Used hierarchical logistic regression model; c-statistic of 0.69, which indicates moderate model discrimination

Panel Member #5: I have a few concerns, and would appreciate if developers could address the following issues:

1. Interpretation of Table 4 (Adjusted OR and 95% CIs for the AMI Mortality Hierarchical Logistic Regression Model over Different Time Periods in the Testing Dataset), especially for factors associated with lower risk of mortality. Could some of these 'protective' factors be due to collinearity with other risk-factors? Were results assessed for clinical plausibility (e.g., Hypertension, Stroke)?
2. I could not identify the results and interpretation of the estimation of average hospital and patient effects related to social risk factors described in section 2b3.3a ("To do this, we performed a decomposition analysis to assess the independent effects of the SRF variables at the patient level and the hospital level.").

3. The decision to not include social risk factors in the model is supported mainly by testing results of no added predictive power and no change in hospital performance rankings. It would be useful to know the rate of hospitals that would have change rank if social-risk factors would have been included, which would provide information on the practical implication not informed by a correlation coefficient between RSRRs for each hospital with and without dual eligibility added. Regarding the result of no added predictive power, have similar considerations been applied to significant clinical factors included in the model, or even more, to non-significant clinical factors which are also expected to have no impact on the model's predictive power and hospital ranking?

Panel Member #6: Very robust risk modeling**.** However, after supplying rationale for including social risk adjustment (SRA) and testing in the model, the developers decide not to include it even though the odds ratios given for their impact in the multivariable model are comparable to other risk factors that are included because they state that there is no impact on the overall c-statistic for the model and the average impact per hospital is negligible. This is not adequate:

1) They do not apply to the same rationale nor test for other risk factors of comparable odds ratios that are included in the model
2) The impact on the overall model or the average impact on hospitals may be small, but the impact on certain hospitals may be great. Unless they check for the net reclassification index or other metric to see how the distribution of ratings would be impacted across the entire spectrum of hospitals rather than averaging the impact they simply do not know.

Panel Member #7: Fair, c-statistic 0.69, unchanged with inclusion of some SES variables. Is there re-arrangement (change in identification of outliers) when SES variables are included (rearrangement under the ROC curve?) May be a moot point because, "the relationship between dual-eligible status and AHRQ low SES is in the opposite direction than what has been the expressed concern of stakeholders interested in adding such adjustment to the models."

Panel Member #8: The methods and results from the risk model are good.

For cost/resource use measures ONLY:

17. **Are the specifications in alignment with the stated measure intent?**

    ☐ **Yes** ☐ **Somewhat** ☐ **No (If "Somewhat" or "No," please explain)**

18. **Describe any concerns of threats to validity related to attribution, the costing approach, carve outs, or truncation (approach to outliers):**

VALIDITY: TESTING

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

    ☒ **Face validity**

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

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

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

    Submission document: Testing attachment, section 2b2.2

    Panel Member #1: The developer used appropriate approach of correlating the measure score with CMS's Hospital Star Rating mortality group score and the overall hospital star rating.

    Panel Member #2: Facility-level correlations with the Star-Rating readmissions score, CMS's Overall Hospital Star Rating, and TJA Surgical Volume.

    Panel Member #3: Compared to star-rating mortality scores – corr coef -0.68

Panel Member #4: For empirical validity testing, compared the hospital's performance on the AMI mortality measure to the hospital's Mortality domain star rating and the hospital's overall Summary star rating. **Concerns with demonstrating validity by using a comparator measure that includes the measure being tested.** (we would expect there to be some correlation!)

Panel Member #5: Face validity was supported during the measure development phase based on national guidelines for publicly reported outcomes measures, and the inclusion of consultation with CMS outside experts and with the public. Empirical testing against other similar measures were appropriate.

Panel Member #6: Comparison with CMS star ratings—not clear that this metric is not included in those ratings, which would make correlation likely by definition. Correlation with alternative source such as National Inpatient Sample might help improve validity.

Panel Member #7: Score correlation with Hospital Star Rating mortality group score and Overall Hospital Star Rating.

Panel Member #8: At the entity level, the measure score was correlated with the CMS's Hospital Star Rating mortality group score, which is derived from this HF measure and other related measures. In a sense, this is checking of the HF measure is related to the latent variable that it was used to construct. It would be indeed surprising and concerning if this hypothesis were not supported. The measure was also correlated with the CMS's Overall Hospital Star Rating, which only indirectly contains the measure through the mortality score. It is reasonably hypothesized that the correlation would be positive but lower than the correlation with the mortality score.

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

    Panel Member #1: The measure score correlation with star-rating mortality groups was moderately strong and in expected direction (-0.676), while the correlation with overall star rating group was weaker but still in the expected direction (-0.114). These findings provide support for the validity of the measures.

    Panel Member #2: Correlation with Star-Rating mortality score is -0.676. Correlation Star-Rating summary score is -0.114. Negative correlation is the desired direction in this case. Correlations with overall star rating is low.

    Panel Member #4: Moderate correlations (-0.409 and -0.204) with Mortality domain star rating and Summary star rating

    Panel Member #5: Empirical testing results are satisfactory, supporting moderate evidence of validity against other related measures.

    Panel Member #6: Weak—graphic representation of box plots without specific correlation statistics

    Panel Member #7: Directions of correlations are as expected.

    Panel Member #8: The correlation between HF RSMRs and the Star-Rating mortality score is -0.676, which suggests that hospitals with lower HF RSMRs are more likely to have higher Star-Rating mortality scores. The correlation between HF RSMRs and Star-Rating summary score is -0.114, which suggests that hospitals with lower HF RSMRs are more likely to have higher Star-Rating summary scores.

23. **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)

24. **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)

25. **OVERALL RATING OF VALIDITY considering 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.)

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

    Panel Member #1: The validity testing results were acceptable, particularly the measure score correlation with mortality star-rating scores.

    Panel Member #3: Risk adjustment model is unlikely to reflect contemporary care because it was developed using data from 1998. This is a major threat to measure validity.

    Panel Member #4: Concerns with the choice of the two measures chosen (Mortality star rating & Summary star rating) to empirically test this measure's validity; a stronger choice would be measures that do not already include the measure under study.

    Panel Member #5: Results suggest low to moderate correlation with similar measures at the score level, thus there is a moderate certainty that the performance measure scores are valid.

    Panel Member #6: On the other hand, if it does, even if the validity testing that they did with the CMS star ratings is weak, the face validity of mortality for patients with heart failure is so unquestionable that no further testing would really be needed. However, testing for >18 population based on 2006 data and not clear that this has been done with this model or with previous model.

    Panel Member #7: Something more than the very, very coarse-grained correlations would be ideal. I am not certain how feasible such an approach would be.

    Panel Member #8: Although the correlation analysis with other similar measures is common, I would prefer to see an analysis of the hypothesized relationships between hospital processes or structures and outcomes. The development, context, and accuracy of the risk model is good.

FOR COMPOSITE MEASURES ONLY: Empirical analyses to support composite construction

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

    ☐ **High**

    ☐ **Moderate**

    ☐ **Low**

    ☐ **Insufficient**

28. **Briefly explain rationale for rating of EMPIRICAL ANALYSES TO SUPPORT COMPOSITE CONSTRUCTION**

ADDITIONAL RECOMMENDATIONS

29. **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.**

Brief Measure Information

NQF #: 0229

Corresponding Measures:

De.2. Measure Title: Hospital 30-Day, All-Cause, Risk-Standardized Mortality Rate (RSMR) Following Heart Failure (HF) Hospitalization

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

De.3. Brief Description of Measure: The measure estimates a hospital-level 30-day, all-cause, risk-standardized mortality rate for patients discharged from the hospital with a principal diagnosis of HF. Mortality is defined as death for any cause within 30 days after the date of admission for the index admission. CMS annually reports the measure for patients who are 65 years or older and enrolled in fee-for-service (FFS) Medicare and hospitalized in non-federal hospitals or are patients hospitalized in Veterans Health Administration (VA) facilities.

1b.1. Developer Rationale: The goal of this measure is to improve patient outcomes by providing patients, physicians, hospitals, and policy makers with information about hospital-level, risk-standardized mortality rates following hospitalization for HF. 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.

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

S.4. Numerator Statement: The outcome for this measure is 30-day all-cause mortality. We define mortality as death from any cause within 30 days from the date of admission for patients 65 and older hospitalized with a principal diagnosis of HF.

S.6. Denominator Statement: This claims-based measure is used for a cohort of patients aged 65 years or older.

The cohort includes admissions for patients aged 65 years and older discharged from the hospital with a principal discharge diagnosis of HF and with a complete claims' history for the 12 months prior to admission. The measure is publicly reported by CMS for those patients 65 years and older who are Medicare FFS or VA beneficiaries admitted to non-federal or VA hospitals, respectively.

Additional details are provided in S.7 Denominator Details.

S.8. Denominator Exclusions: The mortality measures exclude index admissions for patients:

1. Discharged alive on the day of admission or the following day who were not transferred to another acute care facility.
2. With inconsistent or unknown vital status or other unreliable demographic (age and gender) data.
3. Enrolled in the Medicare hospice program or used VA hospice services any time in the 12 months prior to the index admission, including the first day of the index admission.
4. Discharged against medical advice (AMA); or
5. Patients undergoing left ventricular assist device (LVAD) implantation or heart transplantation during an index admission or who have a history of LVAD or heart transplant in the preceding year.

For patients with more than one admission for a given condition each year, only one index admission for that condition is randomly selected for inclusion in the cohort for each year.

De.1. Measure Type: Outcome

S.17. Data Source: Claims, Enrollment Data, Other

S.20. Level of Analysis: Facility

IF Endorsement Maintenance – Original Endorsement Date: May 09, 2007 **Most Recent Endorsement Date:** Feb 19, 2016

IF this measure is included in a composite, NQF Composite#/title:

IF this measure is paired/grouped, NQF#/title:

De.4. IF PAIRED/GROUPED, what is the reason this measure must be reported with other measures to appropriately interpret results? This measure is paired with a measure of hospital-level, all-cause, 30-day, risk-standardized readmission (RSRR) following HF hospitalization.

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

NQF_evidence_HFmortality_Fall2020_final_7.22.20.docx

1a.1 For Maintenance of Endorsement: Is there new evidence about the measure since the last update/submission?

Do not remove any existing information. If there have been any changes to evidence, the Committee will consider the new evidence. Please use the most current version of the evidence attachment (v7.1). Please use red font to indicate updated evidence.

Yes

1a. Evidence (subcriterion 1a)

NATIONAL QUALITY FORUM — Evidence (subcriterion 1a)

Measure Number (*if previously endorsed*)**:** 0229

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

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

Date of Submission: **11/2/2020**

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

Outcome

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

☐ Patient-reported outcome (PRO):

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

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

☐ Process:

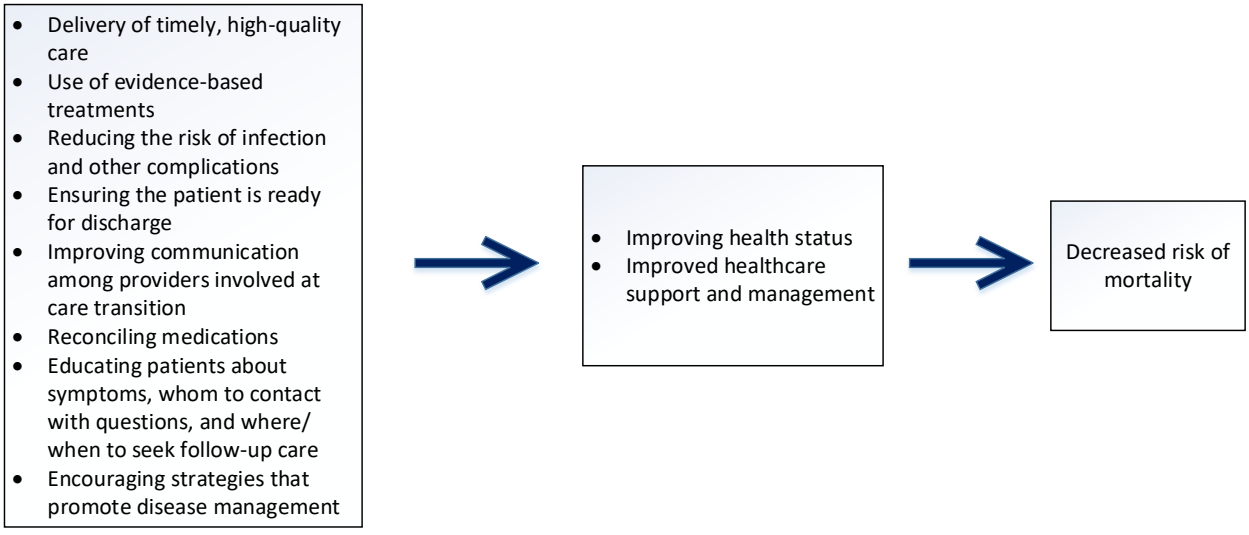
☐ Appropriate use measure:

☐ Structure:

☐ Composite:

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

Figure 1. HF Mortality Logic Model



The goal of this measure is to directly improve patient outcomes by measuring risk-standardized rates of mortality. Measurement of patient outcomes, including mortality, allows for a broad view of quality of care that encompasses more than what can be captured by individual process-of-care measures. As described

below, mortality is likely to be influenced by a broad range of clinical activities such as the prevention of complications and the provision of evidenced-based care.

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

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

****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.

Heart failure (HF) prevalence approaches 10 per 1,000 people in the 65 years and older population each year (NHLBI 2007) and is the most common discharge diagnosis among the elderly (Jessup and Brozena 2003). Prevalence of HF in the U.S. is estimated at more than 6 million cases (Mozaffarian 2015, Lloyd-Jones 2009; Jackson 2018; Benjamin 2020) and is suspected to be the leading cause of death in people over age 65. The lifetime risk of HF is estimated at 1 in 5 at 40 years of age, and the prevalence in the aging US population is expected to increase by 46% by 2030 (Heidenreich 2013). Total direct medical costs of HF were estimated at $30.7 billion in 2012 and are projected to increase by about 127% to $69.7 billion by 2030 (Jackson 2018; Heidenreich 2013).

According to the 2015 AHA (American Hospital Association) update report, one in nine deaths has HF mentioned on the death certificate. In 2011, HF any-mention mortality on death certificates was 284,388, and HF was determined to be the underlying cause in 58,309 of those deaths in 2011 (Mozaffarian 2015, National Center for Health Statistics 2011). There are about 870,000 new HF cases annually (Mozaffarian 2015). Survival after HF diagnosis has improved over time; however, the death rate remains high with about 50% of people diagnosed with HF dying within 5 years (Mozaffarian 2015, Levy et al. 2002, Roger et al. 2004, Jackson et al. 2018). Among Medicare beneficiaries, the overall one-year HF mortality rate declined slightly from 1998 to 2008 but remained high at 29.6% of the population (Chen et al. 2011). Rates of mortality decline were uneven across states.

Clinical experience suggests that the care for these patients is highly variable, and studies suggest quality gaps in hospital care—particularly in the transition to outpatient care (Albert 2009, Jha 2005; Patel et al., 2018). Moreover, there is substantial inter-hospital variation in the risk of death that is not clearly explained by differences in case mix (Lahewala et al., 2018; Roshanghalb et al., 2019; Desai et al., 2018). 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 HF mortality measure is thus intended to inform quality-of-care improvement efforts, as individual process-based performance measures cannot encompass all the complex and critical aspects of care within a

hospital that contribute to patient outcomes. Many stakeholders, including patient organizations, are interested in outcomes measures that allow patients and providers to assess relative outcomes performance for hospitals.

The diagram above indicates some of the many care processes that can influence mortality risk. Numerous studies have demonstrated that appropriate and timely treatment for HF patients can reduce the risk of mortality within 30 days of hospital admission (Hunt 2009, Jha 2007; Kao 2016). Other studies have highlighted how coordinated care for HF patients has been effective in reducing all-cause mortality after HF admission. For instance, the Project RED intervention, an intervention focused on reinforcing coordination of follow-up appointments and testing, medication reconciliation, patient discharge planning, patient education, and post-admission services and durable medical equipment, has showed promise in reducing mortality and costs of care for HF patients (Patel et al., 2018). Additionally, trials of interventions which improve patient education upon discharge have been shown to improve survival for HF patients (McAllister 2001). Evidence that hospitals have been able to reduce mortality rates through these quality-of-care initiatives illustrates the degree to which hospital practices can affect mortality rates.

References

Albert NM, Yancy CW, Liang L, et al. Use of aldosterone antagonists in heart failure. *JAMA.* 2009;302(15):1658-1665.

Benjamin EJ, Muntner P, Alonso A, et al. Heart Disease and Stroke Statistics-2019 Update: A Report From the American Heart Association [published correction appears in Circulation. 2020 Jan 14;141(2): e33]. *Circulation*. 2019;139(10): e56–e528. doi:10.1161/CIR.0000000000000659.

Chen J, Normand SL, Wang Y, Krumholz HM. National and regional trends in heart failure hospitalization and mortality rates for Medicare beneficiaries, 1998–2008. *JAMA*. 2011; 306:1669–1678.

Desai NR, Ott LS, George EJ, et al. Variation in and Hospital Characteristics Associated With the Value of Care for Medicare Beneficiaries With Acute Myocardial Infarction, Heart Failure, and Pneumonia. *JAMA Netw Open*. 2018;1(6): e183519. Published 2018 Oct 5. doi:10.1001/jamanetworkopen.2018.3519.

Heidenreich PA, Albert NM, Allen LA, et al. Forecasting the impact of heart failure in the United States: a policy statement from the American Heart Association. Circ Heart Fail. 2013;6(3):606–619. doi:10.1161/HHF.0b013e318291329a.

Hunt SA, Abraham WT, Chin MH, Feldman AM, Francis GS, Ganiats TG, Jessup M, Konstam MA, Mancini DM, Michl K, Oates JA, Rahko PS, Silver MA, Stevenson LW, Yancy CW; American College of Cardiology Foundation; American Heart Association. 2009 Focused update incorporated into the ACC/AHA 2005 Guidelines for the Diagnosis and Management of Heart Failure in Adults A Report of the American College of Cardiology Foundation/American Heart Association Task Force on Practice Guidelines Developed in Collaboration With the International Society for Heart and Lung Transplantation. J Am Coll Cardiol. 2009 Apr 14;53(15): e1-e90.

Jackson SL, Tong X, King RJ, Loustalot F, Hong Y, Ritchey MD. National Burden of Heart Failure Events in the United States, 2006 to 2014. *Circ Heart Fail.* 2018;11(12): e004873. doi:10.1161/CIRCHEARTFAILURE.117.004873.

Jessup M, Brozena S. Medical progress: heart failure. N Engl J Med 2003; 348:2007–18.

Jha AK, Li Z, Orav EJ, Epstein AM. Care in U.S. Hospitals — The Hospital Quality Alliance Program. *New England Journal of Medicine*. 2005;353(3):265-274.

Jha AK, Orav EJ, Li Z, Epstein AM. The inverse relationship between mortality rates and performance in the Hospital Quality Alliance measures. Health Aff (Millwood) 2007 Jul-Aug;26(4):1104-10.

Kao, D.P., J. Lindenfeld, D. Macaulay, H.G. Birnbaum, J.L. Jarvis, U.S. Desai, and R.L. Page, 2nd, Impact of a Telehealth and Care Management Program on All-Cause Mortality and Healthcare Utilization in Patients with Heart Failure. Telemed J E Health, 2016. 22(1): p. 2-11.

Lahewala S, Arora S, Tripathi B, et al. Heart failure: Same-hospital vs. different-hospital readmission outcomes [published correction appears in Int J Cardiol. 2020 Jun 15; 309:100]. *Int J Cardiol*. 2019; 278:186-191. doi: 10.1016/j.ijcard.2018.12.043.

Levy D, Kenchaiah S, Larson MG, Benjamin EJ, Kupka MJ, Ho KK, Murabito JM, Vasan RS. Long-term trends in the incidence of and survival with heart failure. *N Engl J Med*. 2002; 347:1397–1402.

Lloyd-Jones D et al, American Heart Association Statistics Committee and Stroke Statistics Subcommittee. Heart disease and stroke statistics--2010 update: a report from the American Heart Association. Circulation. 2010 Feb 23;121(7): e46-e215. Epub 2009 Dec 17.

McAllister FA, Lawson FME, Teo KK, Armstrong PW: A systematic review of randomized trials of disease management programs in heart failure. Am J Med 2001, 110:378-384

Mozaffarian D, Benjamin EJ, Go AS, et al. Heart disease and stroke statistics--2015 update: a report from the American Heart Association. *Circulation.* 2015;131(4): e29-322.

National Center for Health Statistics. Mortality Multiple Cause Micro-data Files, 2011. Public-use data file and documentation. NHLBI tabulations. http://www.cdc.gov/nchs/data_access/Vitalstatsonline.htm#Mortality_Multiple. Accessed July 3, 2014.

National Heart, Lung, and Blood Institute. Unpublished tabulation of NHANES, 1971-1975, 1976-1980, 1988-1994, 1999-2002, 2003-2006, and extrapolation to the U.S. population, 2007.

Pandey A, Patel KV, Liang L, et al. Association of Hospital Performance Based on 30-Day Risk-Standardized Mortality Rate With Long-term Survival After Heart Failure Hospitalization: An Analysis of the Get With The Guidelines-Heart Failure Registry. *JAMA Cardiol*. 2018;3(6):489-497. doi:10.1001/jamacardio.2018.0579.

Patel PH, Dickerson KW. Impact of the Implementation of Project Re-Engineered Discharge for Heart Failure patients at a Veterans Affairs Hospital at the Central Arkansas Veterans Healthcare System. *Hosp Pharm*. 2018;53(4):266-271. doi:10.1177/0018578717749925.

Roger VL, Weston SA, Redfield MM, Hellermann-Homan JP, Killian J, Yawn BP, Jacobsen SJ. Trends in heart failure incidence and survival in a community-based population. *JAMA*. 2004; 292:344–350.

Roshanghalb A, Mazzali C, Lettieri E. Multi-level models for heart failure patients' 30-day mortality and readmission rates: the relation between patient and hospital factors in administrative data. *BMC Health Serv Res*. 2019;19(1):1012. Published 2019 Dec 30. doi:10.1186/s12913-019-4818-2.

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

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

☐ Clinical Practice Guideline recommendation (with evidence review)

☐ US Preventive Services Task Force Recommendation

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

☐ Other

Systematic Review	Evidence
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 (SYSTEMATIC REVIEW).	*
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?	*

*cell intentionally left blank

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 HF. 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.

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

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

Variation in mortality rates indicates opportunity for improvement. We conducted analyses using data from July 1, 2016 to June 30, 2019 Medicare claims and VA administrative data (n= 1,081,897 admissions from 4,637 hospitals).

The three-year hospital-level risk-standardized mortality rates (RSMRs) have a mean of 11.4% and range from 5.3-18.5% in the study cohort. As shown below, the median risk-standardized rate is 11.4%. The distribution of RSMRs across hospitals is shown below:

Distribution of Hospital HF RSMRs over Different Time Periods

Results for each data year

Characteristic//07/2016-06/2017//07/2017-06/2018//07/2018-06/2019//07/2016-06/2019

Number of Hospitals// 4,530// 4,504// 4,484// 4,637

Number of Admissions// 353,028// 365,354// 363,515// 1,081,897

Mean (SD (Standard Deviation))// 11.7(1.2)// 11.4(1.1)// 11.1(1.2)// 11.4(1.6)

Range (Min-Max)// 6.5-17// 7.5-18// 5.7-17// 5.3-18.5

Minimum// 6.5// 7.5// 5.7// 5.3

10th percentile//10.3//10.1//9.6//9.5

20th percentile//10.9//10.7//10.2//10.2

30th percentile//11.2//11.0//10.6//10.7

40th percentile//11.4//11.2//10.8//11.1

50th percentile//11.6//11.3//10.9//11.4

60th percentile//11.9//11.6//11.2//11.7

70th percentile//12.1//11.8//11.5//12.1

80th percentile//12.5//12.2//11.9//12.6

90th percentile//13.1//12.8//12.6//13.4

Maximum//17.0//18.0//17.0//18.5

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

N/A

1b.4. Provide disparities data from the measure as specified (current and over time) by population group, e.g., by race/ethnicity, gender, age, insurance status, socioeconomic status, and/or disability. (This is required for maintenance of endorsement. *Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included.) For measures that show high levels of performance, i.e., "topped out," disparities data may demonstrate an opportunity for improvement/gap in care for certain sub-populations. This information also will be used to address the sub-criterion on improvement (4b1) under Usability and Use.*

Distribution of 30-day HF RSMRs by Proportion of Dual Eligible Patients:

Data Source: Medicare FFS claims, VA data, and Master Beneficiary Summary File (MBSF) data

Dates of Data: July 2016 through June 2019

Variation in RSMRs across hospitals (with at least 25 cases) by proportion of patients with social risk//

Description of Social Risk Variable//Dual Eligibility

Quartile//Q1//Q4

Social Risk Proportion (%)// (0-8.37)// (34.43-100)

# of Hospitals//910//910

100% Max//17.7//18.5

90%//13.6//13.4

75%//12.7//12.3

50%//11.5//11.2

25%//10.3//10.1

10%//9.4//9.2

0% Min//6.7//5.3

Distribution of 30-day HF RSMRs by Proportion of Patients with AHRQ SES Index Scores:

Data Source: Medicare FFS claims, VA data, and the American Community Survey (ACS) data

Dates of Data: July 2016 through June 2019 (claims); 2013-2017 (ACS)

Variation in RSMRs across hospitals (with at least 25 cases) by proportion of patients in lower and upper social risk quartiles//

Description of Social Risk Variable //AHRQ SES Index

Quartile//Q1//Q4

Social Risk Proportion (%)// (0-10.24)// (23.59-100)

# of Hospitals//911//911

100% Max//17.1//17.1

90%//13.3//13.4

75%//12.3//12.3

50%//11.1//11.1

25%//10.0//10.0

10%//9.0//9.0

0% Min//6.1//5.3

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

N/A

2. Reliability and Validity—Scientific Acceptability of Measure Properties

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

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

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

Cardiovascular, Cardiovascular: Congestive Heart Failure

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

Safety

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

Elderly, Populations at Risk

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

https://www.qualitynet.org/inpatient/measures/mortality

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

This is not an eMeasure Attachment**:**

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

Attachment: NQF_datadictionary_HFmortality_Fall2020_final_7.22.20.xlsx

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

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

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

Not an instrument-based measure

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

No

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

Updates consisted of updating the specifications to include new and modified ICD-10 CM/PCS codes.

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

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

The outcome for this measure is 30-day all-cause mortality. We define mortality as death from any cause within 30 days from the date of admission for patients 65 and older hospitalized with a principal diagnosis of HF.

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

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

The measure counts all deaths (including in-hospital deaths) for any cause within 30 days of the date of the index HF admission.

Identifying deaths in the FFS measure

As currently reported, we identify deaths for FFS Medicare patients 65 years and older in the Medicare Enrollment Database (EDB) and for VA patients in the VA data.

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

This claims-based measure is used for a cohort of patients aged 65 years or older.

The cohort includes admissions for patients aged 65 years and older discharged from the hospital with a principal discharge diagnosis of HF and with a complete claims' history for the 12 months prior to admission. The measure is publicly reported by CMS for those patients 65 years and older who are Medicare FFS or VA beneficiaries admitted to non-federal or VA hospitals, respectively.

Additional details are provided in S.7 Denominator Details.

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

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

To be included in the measure cohort used in public reporting, patients must meet the following inclusion criteria:

1. Principal discharge diagnosis of heart failure
2. Enrolled in Medicare fee-for-service (FFS) Part A and Part B for the 12 months prior to the date of the index admission and Part A during the index admission, or those who are VA beneficiaries
3. Aged 65 or over
4. Not transferred from another acute care facility

We have explicitly tested the measure for those aged 65+ years and those aged 65+ years (see Testing Attachment for details).

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

The mortality measures exclude index admissions for patients:

1. Discharged alive on the day of admission or the following day who were not transferred to another acute care facility.
2. With inconsistent or unknown vital status or other unreliable demographic (age and gender) data.
3. Enrolled in the Medicare hospice program or used VA hospice services any time in the 12 months prior to the index admission, including the first day of the index admission.
4. Discharged against medical advice (AMA); or
5. Patients undergoing left ventricular assist device (LVAD) implantation or heart transplantation during an index admission or who have a history of LVAD or heart transplant in the preceding year.

For patients with more than one admission for a given condition each year, only one index admission for that condition is randomly selected for inclusion in the cohort for each year.

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

1. The discharge disposition indicator is used to identify patients alive at discharge. Transfers are identified in the claims when a patient with a qualifying admission is discharged from an acute care hospital and admitted to another acute care hospital on the same day or next day. Patient length of stay and condition is identified from the admission claim.

Rationale: This exclusion prevents inclusion of patients who likely did not have clinically significant HF.

2. Inconsistent vital status or unreliable data are identified if any of the following conditions are met
    1) the patient's age is greater than 115 years:
    2) if the discharge date for a hospitalization is before the admission date;
    3) if the patient has a sex other than 'male' or 'female.'

Rationale: Reliable and consistent data are necessary for valid calculation of the measure.

3. Hospice enrollment in the 12 months prior to or on the index admission is identified using hospice data and the Inpatient standard analytic file (SAF).

Rationale: These patients are likely continuing to seek comfort measures only; thus, mortality is not necessarily an adverse outcome or signal of poor-quality care.

4. Discharges against medical advice (AMA) are identified using the discharge disposition indicator in claims data.

Rationale: Providers did not have the opportunity to deliver full care and prepare the patient for discharge.

5. Patients with LVAD implantation or heart transplantation during an index admission or in the previous 12 months are identified by the corresponding codes for these procedures included in claims data.

Rationale: Patients undergoing implantation of an LVAD designed to offer intermediate to long-term support (weeks to years) as a bridge to heart transplant or destination therapy represent a clinically distinct, highly selected group of patients cared for at highly specialized medical centers.

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

N/A

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

Statistical risk model

If other:

S.12. Type of score:

Rate/proportion

If other:

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

Better quality = Lower score

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

The measure estimates hospital-level 30-day all-cause RSMRs following hospitalization for HF using hierarchical logistic regression models. In brief, the approach simultaneously models data at the patient and hospital levels to account for variance in patient outcomes within and between hospitals [Normand and Shahian, 2007]. At the patient level, it models the log-odds of mortality within 30 days of index admission using age, sex, selected clinical covariates, and a hospital-specific intercept. At the hospital level, it models the hospital-specific intercepts as arising from a normal distribution. The hospital intercept represents the underlying risk of a mortality at the hospital, after accounting for patient risk. The hospital-specific intercepts are given a distribution to account for the clustering (non-independence) of patients within the same hospital. If there were no differences among hospitals, then after adjusting for patient risk, the hospital intercepts should be identical across all hospitals.

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

The "predicted" number of deaths (the numerator) is calculated by using the coefficients estimated by regressing the risk factors and the hospital-specific intercept on the risk of mortality. The estimated hospital-specific intercept is added coefficients multiplied by the patient characteristics. The results are transformed and summed over all patients attributed to a hospital to get a predicted value. The "expected" number of deaths (the denominator) is obtained in the same manner, but a common intercept using all hospitals in our sample is added in place of the hospital-specific intercept. The results are transformed and summed over all patients in the hospital to get an expected value. To assess hospital performance for each reporting period, we re-estimate the model coefficients using the years of data in that period.

This calculation transforms the ratio of predicted over expected into a rate that is compared to the national observed readmission rate. The hierarchical logistic regression models are described fully in the original methodology report posted on QualityNet [https://qualitynet.org/inpatient/measures/mortality/methodology].

References:

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

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

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

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

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

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

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

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

If other, please describe in S.18.

Claims, Enrollment Data, Other

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

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

Data sources for the Medicare FFS measure:

Medicare Part A Inpatient and Part B Outpatient Claims: This data source contains claims data for FFS inpatient and outpatient services including Medicare inpatient hospital care, outpatient hospital services, skilled nursing facility care, some home health agency services, as well as inpatient and outpatient physician claims for the 12 months prior to an index admission.

Medicare Enrollment Database (EDB): This database contains Medicare beneficiary demographic, benefit/coverage, and vital status information. This data source was used to obtain information on several inclusion/exclusion indicators such as Medicare status on admission as well as vital status. These data have previously been shown to accurately reflect patient vital status (Fleming et al., 1992). The Master Beneficiary Summary File (MBSF) is an annually created file derived the EDB that contains enrollment information for all Medicare beneficiaries including dual eligible status. Years 2016-2019 were used.

Veterans' Health Administration (VA) Data: This data source contains data for VA inpatient and outpatient services including inpatient hospital care, outpatient hospital services, skilled nursing facility care, some home health agency services, as well as inpatient and outpatient physician data for the 12 months prior to and including each index admission. Unlike Medicare FFS patients, VA patients are not required to have been enrolled in Part A and Part B Medicare for the 12 months prior to the date of admission.

The American Community Survey (2013-2017): The American Community Survey data is collected annually, and an aggregated 5-years data were used to calculate the Agency for Healthcare Research and Quality (AHRQ) Socioeconomic Status (SES) composite index score.

References:

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

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

No data collection instrument provided

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

Facility

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

Inpatient/Hospital

If other:

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

N/A

2. Validity – See attached Measure Testing Submission Form

NQF_testing_HFmortality_Fall2020_final_10.27.20_updated_11_17_20.docx

2.1 For maintenance of endorsement

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

Yes

2.2 For maintenance of endorsement

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

Yes

2.3 For maintenance of endorsement

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

Yes - Updated information is included

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

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

Measure Number (*if previously endorsed*)**:** 0229
Measure Title: Hospital 30-day, all-cause, risk-standardized mortality rate (RSMR) following heart failure (HF) hospitalization
Date of Submission: **8/3/2020**

Type of Measure:

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

*cell intentionally left blank

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

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

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

Measure Specified to Use Data From: (*must be consistent with data sources entered in S.17*)	Measure Tested with Data From:
☐ abstracted from paper record	☐ abstracted from paper record
☒ claims	☒ claims
☐ registry	☐ registry
☐ abstracted from electronic health record	☐ abstracted from electronic health record
☐ eMeasure (HQMF) implemented in EHRs	☐ eMeasure (HQMF) implemented in EHRs
☒ other: Medicare Enrollment Data (including the Master Beneficiary Summary File), VHA Administrative Data	☒ other: Census Data/American Community Survey, VHA Administrative Data, Medicare Enrollment Data (including the Master Beneficiary Summary File)

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 data used for testing included Medicare Parts A and B claims as well as the Medicare Enrollment Database (EDB). Additionally, census as well as enrollment data were used to assess socioeconomic factors (dual eligible variable obtained through enrollment data; Agency for Healthcare Research and Quality [AHRQ] socioeconomic status [SES] index obtained through census data). Veterans' Health Administration (VHA) data are also included in the testing dataset. The dataset used varies by testing type; see Section 1.7 for details.

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

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

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

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

For this measure, hospitals are the measured entities. All non-federal, short-term acute care inpatient US hospitals (including territories) with Medicare fee-for-service (FFS) beneficiaries aged 65 years or over are included. In addition, for the testing period presented, VA hospitals and their patients 65 years and older are included in the measure. The number of measured entities (hospitals) varies by testing type; see Section 1.7 for details.

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

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

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

The datasets, dates, number of measured hospitals, and number of admissions used in each type of testing are in Table 1.

Measure Development

For measure development, we used Medicare administrative claims data (1998). The dataset also included administrative data on each patient for the 12 months prior to the index admission. The dataset contained inpatient and facility outpatient claims and Medicare enrollment database (EDB) data. We randomly split the data into two equal samples: the Development Dataset and Internal Validation Dataset.

Measure Testing

For analytical updates for this measure, we used three years of Medicare administrative claims data (July 2016 – June 2019). The dataset also included administrative data on each patient for the 12 months prior to the index admission. The dataset contained inpatient and facility outpatient claims and Medicare enrollment database (EDB) data. The dataset also included administrative data from the VHA as these hospitals are currently publicly reported for this measure.

Table 1. Dataset Descriptions

Dataset	Applicable Section in the Testing Attachment	Description of Dataset
Development and Validation Datasets (Medicare Fee-For-Service Administrative Claims Data)	Section 2b3 Risk Adjustment/Stratification **2b3.6. Statistical Risk Model Discrimination Statistics** **2b3.7. Statistical Risk Model Calibration Statistics**	Entire Cohort: Dates of Data: 1998 Number of admissions = 444,581 Number of measured hospitals: 5,088 This cohort was randomly split for initial model testing. First half of split sample -Number of Admissions: 222,424 -Number of Measured Hospitals: 5,087 Second half of split sample -Number of Admissions: 222,157 -Number of Measured Hospitals: 5,088
Testing Dataset (Medicare Fee-For-Service Administrative Claims Data (July 1, 2016 – June 30, 2019)	Section 2a2 Reliability Testing Section 2b1 Validity Testing Section 2b2 Testing of Measure Exclusion Section 2b3 Risk Adjustment/Stratification **Section 2b3.6. Statistical Risk Model Discrimination Statistics** Section 2b4 Meaningful Differences	Dates of Data: July 2016 – June 2019 Number of admissions = 1,081,897 Patient Descriptive Characteristics: mean age = 80.7 years; % male = 48.3 Number of measured hospitals: 4,637
The American Community Survey (ACS)	Section 2b3: Risk adjustment/Stratification for Outcome or Resource Use Measures	Dates of Data: 2013-2017 We used the AHRQ SES index score derived from the American Community Survey (2013-2017) to study the association between the 30-day mortality outcome and SRFs. The AHRQ SES index score is based on beneficiary 9-digit zip code level of residence and incorporates 7 census variables found in the American Community Survey.

Dataset	Applicable Section in the Testing Attachment	Description of Dataset
Master Beneficiary Summary File (MBSF)	Section 2b3: Risk adjustment/Stratification for Outcome or Resource Use Measures	Dates of Data: July 2016 – June 2019 We used dual eligible status (for Medicare and Medicaid) derived from the MBSF to study the association between the 30-day measure outcome and dual-eligible status.

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

We selected social risk factor (SRF) variables to analyze after reviewing the literature and examining available national data sources. We sought to find variables that are consistently captured in a reliable fashion for all patients in this measure. There is a large body of literature linking various SRFs to worse health status and higher mortality over a lifetime. Income, education, and occupation are the most examined SRFs studied. The causal pathways for SRF variable selection are described below in Section 2b3.3a. Unfortunately, these variables are not available at the patient-level for this measure. Therefore, proxy measures of income, education level and economic status were selected.

The SRF variables used for analysis were:

- Dual eligible status: Dual eligible status (i.e., enrolled in both Medicare and Medicaid) patient-level data is obtained from the CMS Master Beneficiary Summary File (MBSF).

  Following guidance from ASPE and a body of literature demonstrating differential health care and health outcomes among dual eligible patients, we identified dual eligibility as a key variable (ASPE 2016; ASPE 2020). 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 outcome. However, the threshold for over 65-year-old Medicare patients is valuable, as it considers both income and assets and is consistently applied across states for the older population. We acknowledge that it is important to test a wider variety of SRFs including key variables such as education and poverty level; therefore, we also tested a validated composite based on census data linked to as small a geographic unit as possible.

- AHRQ-validated SES index score (summarizing the information from the following seven 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.

  Finally, we selected the AHRQ SES index score because it is a well-validated variable that describes the average SES of people living in defined geographic areas (Bonito et al., 2008). 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 considered the area deprivation index (ADI) among many other potential indicators when we initially evaluated the impact of SDS indicators. We ultimately did not include the ADI at the time, partly due to the fact that the coefficients used to derive ADI had not been updated for many years. Recently, the coefficients for ADI have been updated and therefore we compared the ADI with the AHRQ SES Index and found them to be highly correlated. In this submission, we present analyses using the census block level, the most granular level possible using American Community Survey (ACS) data. A census block group is a geographical unit used by the US Census Bureau which is between the

census tract and the census block. It is the smallest geographical unit for which the bureau publishes sample data. The target size for block groups is 1,500 and they typically have a population of 600 to 3,000 people. We used 2013-2017 ACS data and mapped patients' 9-digit ZIP codes via vendor software to the census block group level. Given the variation in cost of living across the country, the median income and median property value components of the AHRQ SES Index were adjusted by regional price parity values published by the Bureau of Economic Analysis (BEA). This provides a better marker of low SES neighborhoods in high expense geographic areas. We then calculated an AHRQ SES Index score for census block groups that can be linked to 9-digit ZIP codes. We used the percentage of patients with an AHRQ SES index score equal to or below 42.7 to define the lowest quartile of the AHRQ SES Index.

References:

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

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

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

Department of Health and Human Services, Office of the Assistant Secretary of Planning and Evaluation (ASPE). 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, 2019.

Department of Health and Human Services, Office of the Assistant Secretary of Planning and Evaluation (ASPE). Second Report to Congress: Social Risk Factors and Performance in Medicare's Value-based Purchasing Programs. 2020; https://aspe.hhs.gov/system/files/pdf/263676/Social-Risk-in-Medicare%E2%80%99s-VBP-2nd-Report.pdf. Accessed July 2, 2020.

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.

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 Aff (Millwood). Aug 2014; 33(8):1314-22.

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

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

Kosar CM, Loomer L, Ferdows NB, Trivedi AN, Panagiotou OA, Rahman M. Assessment of Rural-Urban Differences in Postacute Care Utilization and Outcomes Among Older US Adults. *JAMA Netw Open*. 2020;3(1): e1918738. Published 2020 Jan 3. doi:10.1001/jamanetworkopen.2019.18738.

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

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

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

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

We performed two types of reliability testing. First, we estimated the overall measure score reliability by calculating the intra-class correlation coefficient (ICC) using a split sample (i.e., test-retest) method. Second, we estimated the facility-level reliability (signal-to-noise reliability).

Split-Sample 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. Accordingly, our approach to assessing reliability is to consider the extent to which assessments of a hospital using different but randomly selected subsets of patients produce similar measures of hospital performance. That is, we take a "test-retest" approach in which hospital performance is measured once using a random subset of patients, and then measured again using a second random subset exclusive of the first, and the agreement of the two resulting performance measures compared across hospitals (Rousson, Gasser, and Seifert, 2002).

For split-sample reliability of the measure in aged 65 years and older, we randomly sampled half of patients within each hospital for a three-year period, calculated the measure for each hospital, and repeated the calculation using the second half. Thus, each hospital is measured twice, but each measurement is made using an entirely distinct set of patients. To the extent that the calculated measures of these two subsets agree, we have evidence that the measure is assessing an attribute of the hospital, not of the patients. As a metric of agreement, we calculated the intra-class correlation coefficient (Shrout & Fleiss, 1979), and assessed the values according to conventional standards (Landis & Koch, 1977). Specifically, we used a combined 2016-2019 sample, randomly split it into two approximately equal subsets of patients and calculated the RSMR for each hospital for each sample. The agreement of the two RSMRs was quantified for hospitals in each sample using the intra-class correlation as defined by ICC (2,1) (Shrout & Fleiss, 1979).

Using two non-overlapping random samples provides a conservative estimate of the measure's reliability, compared with using two random but potentially overlapping samples which would exaggerate the agreement. Moreover, because our final measure is derived using hierarchical logistic regression, and a known property of hierarchical logistic regression models is that smaller volume hospitals contribute less 'signal', a split sample using a single measurement period would introduce extra noise, potentially underestimating the actual test-retest reliability that would be achieved if the measure were reported using three years of data. This leads to an underestimate in the actual test-retest reliability that would be achieved if the measure were reported using the full measurement period, as evidenced by the Spearman Brown prophecy formula (Spearman 1910, Brown 1910). We used this formula to estimate the reliability of the measure if the whole cohort were used, based on an estimate from half the cohort.

Signal-to-Noise

We estimated the signal to noise reliability (facility-level reliability), which is the reliability with which individual units (hospitals) are measured. While test re-test reliability is the most relevant metric from the perspective of overall measure reliability, it is also meaningful to consider the separate notion of "unit" reliability, that is, the reliability with which individual units (here, hospitals) are measured. The reliability of any one facility's measure score will vary depending on the number of patients admitted for HF. Facilities with more volume (i.e., with more patients) will tend to have more reliable scores, while facilities with less volume will tend to have less reliable scores. Therefore, we used the formula presented by Adams and colleagues (2010) to calculate facility-level reliability.

Where facility-to-facility variance is estimated from the hierarchical logistic regression model, n is equal to each facility's observed case size, and the facility error variance is estimated using the variance of the logistic distribution (π^2/3). The facility-level reliability testing is limited to facilities with at least 25 admissions for public reporting.

Signal to noise reliability scores can range from 0 to 1. A reliability of zero implies that all the variability in a measure is attributable to measurement error. A reliability of one implies that all the variability is attributable to real difference in performance.

Additional Information

In constructing the measure, we aim to utilize only those data elements from the claims that have both face validity and reliability. We avoid the use of fields that are thought to be coded inconsistently across providers. Specifically, we use fields that are consequential for payment and which are audited. We identify such variables through empiric analyses and our understanding of CMS auditing and billing policies and seek to avoid variables which do not meet this standard.

In addition, CMS has in place several hospital auditing programs used to assess overall claims code accuracy, to ensure appropriate billing, and for overpayment recoupment. CMS routinely conducts data analysis to identify potential problem areas and detect fraud, and audits important data fields used in our measures, including diagnosis and procedure codes and other elements that are consequential to payment.

Furthermore, we assessed the variation in the frequency of the variables over time. Detailed information is presented in the measure's 2020 Condition-Specific Measure Updates and Specifications Report cited below.

References

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

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

Debuhr J, McDowell K, Grady J et al., 2020 Condition-Specific Measure Updates and Specifications Report Hospital-Level 30-Day Risk-Standardized Mortality Measures - Available at: https://www.qualitynet.org/inpatient/measures/mortality/methodology.

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

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

Shrout P, Fleiss J. Intraclass correlations uses in assessing rater reliability. Psychological Bulletin, 1979, 86, 420-3428.

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

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

Measure Score Reliability Results

Split-Sample Reliability

In total, 1,081,887 admissions were included in the analysis, using 3 years of data. After randomly splitting the sample into two halves, there were 539,795 admissions from 4,590 hospitals in one half and 542,102 admissions from 4,637 hospitals in the other half. As a metric of agreement, we calculated the ICC for hospitals with 25 admissions or more. Using the Spearman-Brown prediction formula, the agreement between the two independent assessments of the RSMR for each hospital was 0.632.

Signal-to-Noise

We calculated the signal-to-noise reliability score for each hospital with at least 25 admissions* (see Table 2 below). The median reliability score was 0.79, ranging from 0.34 to 0.99. The 25th and 75th percentiles were 0.58 and 0.90, respectively. The median reliability score demonstrates moderate reliability.

Table 2. Signal-to-noise reliability distribution for HF mortality

Mean	Std. Dev.	Min	5th Percentile	10th Percentile	25th Percentile	Median	75th Percentile	90th Percentile	95th Percentile	Max
0.73	0.19	0.34	0.39	0.44	0.58	0.79	0.90	0.93	0.95	0.99

*Hospital measure scores are calculated for all hospitals (including those that have fewer than 25 admissions) but only reported for those that have at least 25 admissions to ensure hospital results are reliable.

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?*)

Measure Score Reliability Results

The split-sample reliability score of 0.632, discussed in the previous section, represents the lower bound of estimate of the true measure reliability.

Using the approach used by Adams et. al., we obtained the median signal-to-noise reliability score of 0.79, which demonstrates substantial agreement.

Our interpretation of the results is based on the standards established by Landis and Koch (1977):

- < 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

Taken together, these results indicate that there is substantial reliability in the measure score.

References:

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

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

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

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

Stewards of NQF-endorsed measures going through the re-endorsement process are required to demonstrate external validity testing at the time of maintenance review, or if this is not possible, justify the use of face validity only. To meet this requirement for the HF mortality measure, we identified and assessed the measure's correlation with other measures that target the same domain of quality (e.g., complications, safety, or post-procedure utilization) for the same or similar populations. The goal was to identify if better performance in this measure was related to better performance on other relevant structural or outcomes measures. After literature review and consultations with measures experts in the field, there were very few measures identified that assess the same domains of quality. Given that challenge, we selected the following to use for validity testing.

Hospital Star Rating mortality group score: CMS's Hospital Star Rating mortality group score assesses hospitals' overall performance (expressed on Hospital Compare graphically, as stars) based on a weighted average of group scores from the mortality domain. The mortality group is comprised of the mortality measures that are publicly reported on Hospital Compare, including this HF mortality measure. 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 4,637 Medicare FFS hospitals from July 2019. The full methodology for the Overall Hospital Star Rating can be found at: https://www.qualitynet.org/inpatient/public-reporting/overall-ratings/resources.

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 4,637 Medicare FFS hospitals from July 2019. The full methodology for the Overall Hospital Star Rating can be found at: https://www.qualitynet.org/inpatient/public-reporting/overall-ratings/resources.

We examined the relationship of performance the HF mortality measure scores (RSMR) with each of the external measures of hospital quality. For the external measures, the comparison was against performance within quartiles of the mortality group score, or in the case of Star Ratings, to the Star Rating category (1-5 Stars). We predicted the HF mortality scores would be more strongly associated with the Hospital Star Rating mortality group score than the Overall Star Ratings scores, with lower RSMRs associated with better Star Ratings.

Clinical Validity

During original measure development we validated the HF mortality administrative model against a medical record model in the same cohort of patients for which hospital-level HF mortality medical record data are available.

We developed a medical record measure to compare with the administrative measure. We developed a measure cohort with the medical record data using the inclusion/exclusion criteria and risk-adjustment strategy that was consistent with the claims-based administrative measure but using chart-based risk adjusters, such as blood pressure, not available in the claims data. We then matched a sample of the same patients in the administrative data for comparison. The matched sample included 46,700 patients. We compared the output of the two measures, the state performance results, in the same group of patients.

For the derivation of the chart-based model, we used cases identified through a Health Care Financing Administration (now CMS) quality initiative, which sampled admissions from FFS Medicare beneficiaries for several clinical conditions, including HF (Jencks et al., 2000). Cases were identified over a 6-month period within each state, plus the District of Columbia and Puerto Rico, during the period April 1, 1998 through October 31, 1999. Based on the principal discharge diagnosis, approximately 800 HF discharges per state were identified, and the corresponding medical records were abstracted by two clinical data abstraction centers. In states with fewer than 900 HF discharges, all cases were used. The abstractors first sorted eligible claims by age, race, sex, and hospital. Then, they systematically sampled cases from a random starting point. Patients must have been enrolled in FFS Medicare. CMS subsequently conducted a re-measurement using the same data collection methodology for 2000 and 2001 discharges (Jencks et al., 2003), and the combined 1998-2001 data, including 73,832 patients, served as the NHF dataset for development of the chart-based model.

From the medical chart-abstracted HF cases, we linked these files to the corresponding administrative data and mortality data from the Medicare enrollment database. Because only patients aged 65 years and older were included, and some data were excluded based on linkage and other factors, a total of 46,700 HF hospitalizations were used in the analysis.

The same coding and transfer rules described in the HF administrative dataset were used in defining the HF chart dataset.

The chart model was derived in the NHF dataset. The derivation sample contained 46,700 cases with an unadjusted 30-day mortality rate of 11.9%. Twenty-eight covariates were included in the final model, with age having the largest impact on risk. While the administrative mortality models explained about 10-12% of the observed variation and had accuracy of 69-71%, the chart model explained 21-22% of the variation and had accuracy of 75-78%. Moreover, the predictive ability of the model is excellent—observed mortality in the lowest estimated decile is 1.8% for 30-day mortality, compared with 42.4% (30-day mortality) in the highest estimated decile, a range of 40.6%.

Validity Indicated by Established Measure Development Guidelines:

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

References:

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

Jencks SF, Cuerdon T, Burwen DR et al. Quality of medical care delivered to Medicare beneficiaries: a profile at state and national levels. JAMA. 2000; 284:1670-1676.

Jencks SF, Huff ED, Cuerdon T. Change in the quality of care delivered to Medicare beneficiaries, 1998-1999 to 2000-2001. JAMA. 2003; 289:305-312.

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

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

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

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

National Quality Forum. National voluntary consensus standards for patient outcomes, first report for phases 1 and 2: A consensus report http://www.qualityforum.org/projects/Patient_Outcome_Measures_Phases1-2.aspx. Accessed August 19, 2010.

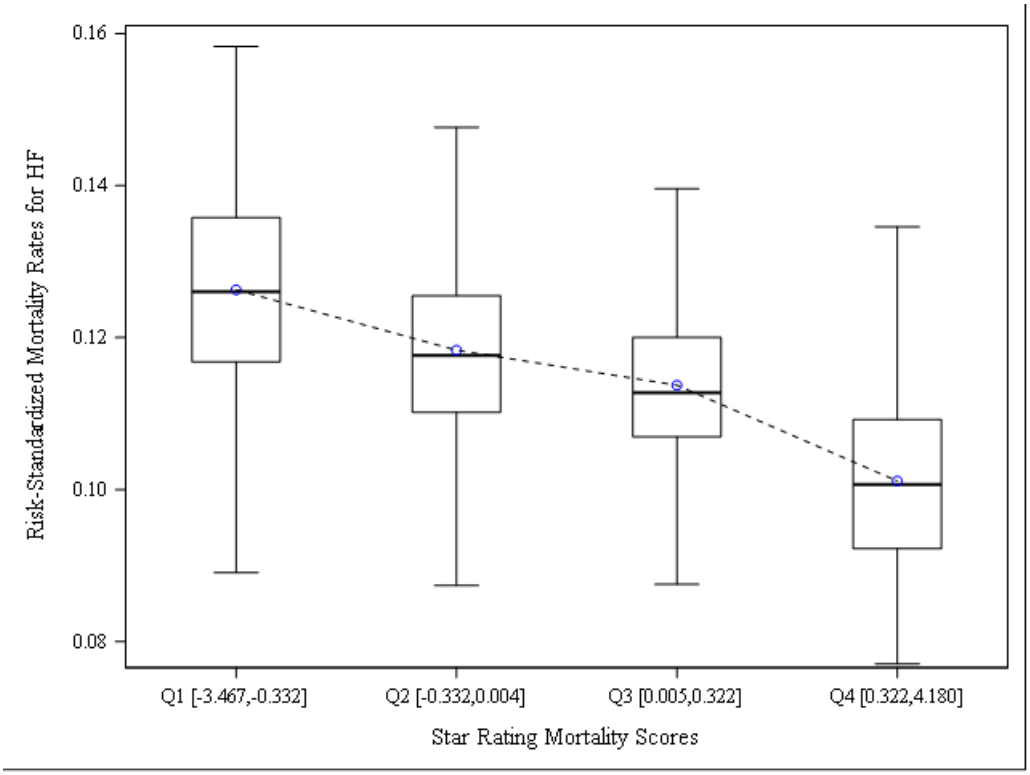
Shahian DM, He X, O'Brien S, et al. Development of a Clinical Registry-Based 30-Day Readmission Measure for Coronary Artery Bypass Grafting Surgery. Circulation 2014; DOI: 0.1161/CIRCULATIONAHA.113.007541. Published online before print June 10, 2014.

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

Comparison to Star-Rating Mortality Scores

Figure 1 shows the box-whisker plots of the HF mortality measure RSMRs within each quartile of Star-Rating mortality scores. The blue circles represent the mean RSMRs of Star-Rating mortality score quartiles. The correlation between HF RSMRs and the Star-Rating mortality score is -0.676, which suggests that hospitals with lower HF RSMRs are more likely to have higher Star-Rating mortality scores.

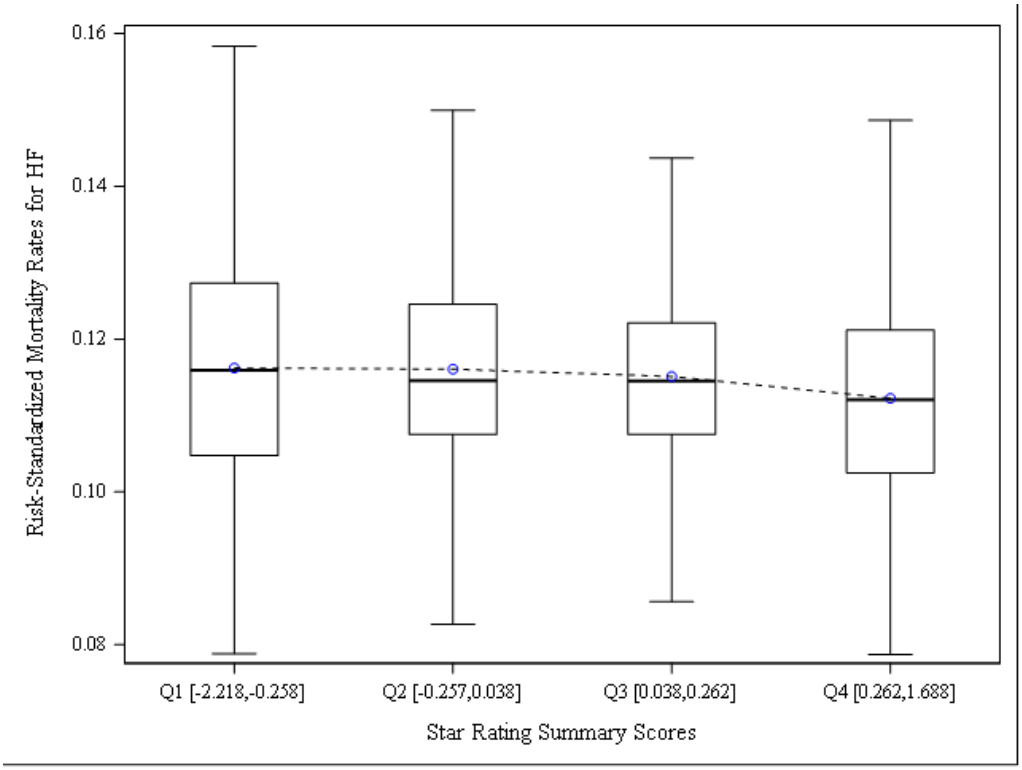
Figure 1 - Box whisker plots of the HF mortality RSMRs within each quartile of Star-rating mortality scores



Comparison to Star-Rating Summary Scores

Figure 2 shows the Box-whisker plots of the HF mortality measure RSMRs within each quartile of Star-Rating summary scores. The blue circles represent the mean RSMRs of Star-Rating summary score quartiles. The correlation between HF RSMRs and Star-Rating summary score is -0.114, which suggests that hospitals with lower HF RSMRs are more likely to have higher Star-Rating summary scores.

Figure 2 - Box whisker plots of the HF mortality RSMRs within each quartile of Star-rating summary scores



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?*)

Empirical Validity Testing

This validation approach compares the 30-day HF mortality measure results against the star rating mortality domain and overall summary scores. Figure 1 and 2 Box Plots results demonstrate an observed trend of lower risk-standardized mortality with higher star ratings score, especially at the extremes, which supports measure score validity. The correlation coefficients associated with the star rating mortality domain scores and the HF mortality measure scores indicate a strong association. A more moderate association is seen with the overall star ratings score, which is to be expected given the measures are calculated by complex statistical models. Overall, the results above show that the trend and direction of this association is in line with what would be expected.

2b2. EXCLUSIONS ANALYSIS

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

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

All exclusions were determined by careful clinical review and have been made based on clinically relevant decisions 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 (**Testing Dataset**). These exclusions are consistent with similar NQF-endorsed outcome measures. Rationales for the exclusions are detailed in data field S.9 (Denominator Exclusions).

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

In the **Testing Dataset (Table 3)**, below is the distribution of exclusions among hospitals with 25 or more admissions:

Table 3. Distribution of exclusions in the testing dataset

Exclusion	N	%	Distribution across hospitals N=3,854 (Min, 25th, 50th, 75th percentile, max)
1. Discharged alive on the day of admission or the following day who were not transferred to another acute care facility	62,467	4.22%	(0.00, 2.25, 4.17,7.14, 35.0)
2. Inconsistent or unknown vital status or other unreliable demographic (age and gender) data	69	0.00%	(0.00, 0.00, 0.00,0.00, 3.13)
3. Enrolled in the Medicare hospice program or used VA hospice services any time in the 12 months prior to the index admission, including the first day of the index admission	20,172	1.36%	(0.00, 0.26, 1.04,1.96, 24.8)
4. Left Ventricular Assist Device (LVAD) or transplant in index admission or prior year	4,672	0.32%	(0.00, 0.00, 0.00,0.14, 8.50)
5. Discharged against medical advice (AMA)	8,859	0.60%	(0.00, 0.00, 0.27,0.82, 9.42)

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

Exclusion 1 (patients who were discharged alive on the day of admission or the following day who were not transferred to another acute care facility) accounts for 4.22% of all index admissions excluded from the initial index cohort. This exclusion represents the majority of all exclusions and is meant to ensure a clinically coherent cohort. This exclusion prevents inclusion of patients who likely did not have clinically significant HF.

Exclusion 2 (patients with inconsistent or unknown vital status or other unreliable demographic [age and gender] data) accounts for less than 0.01% of all index admissions excluded from the initial index cohort. We do not include stays for patients where the age is greater than 115, where the gender is neither male nor female, where the admission date is after the date of death in the Medicare Enrollment Database, or where the date of death occurs before the date of discharge, but the patient was discharged alive.

Exclusion 3 (patients enrolled in the Medicare hospice program or used VA hospice services any time in the 12 months prior to the index admission, including the first day of the index admission) accounts for 1.36% of all index admissions excluded from the initial index cohort. These patients are likely continuing to seek comfort measures only; thus, mortality is not necessarily an adverse outcome or signal of poor-quality care.

Exclusion 4 (patients with LVAD, history of LVAD, transplant, history of transplant) accounts for 0.32% of all index admissions excluded from the initial index cohort. This exclusion is meant to ensure a clinically coherent cohort. Patients undergoing implantation of an LVAD designed to offer intermediate to long-term support

(weeks to years) as a bridge to heart transplant or destination therapy represent a clinically distinct, highly selected group of patients cared for at highly specialized medical centers.

Exclusion 5 (patients who are discharged AMA) accounts for 0.60% of all index admissions excluded from the initial index cohort. This exclusion is needed for acceptability of the measure to hospitals, who do not have the opportunity to deliver full care and prepare the patient for discharge. Given that a very small percentage of patients are being excluded, it is unlikely this exclusion affects the measure score.

After all exclusions are applied, the measure randomly selects one index admission per patient per year for inclusion in the cohort so that each episode of care is mutually independent with a similar probability of the outcome. For each patient, the probability of death changes with each subsequent admission, and therefore, the episodes of care are not mutually independent. Similarly, for the three-year combined data, when index admissions occur during the transition between measure reporting periods (June and July of each year) and both are randomly selected for inclusion in the measure, the measure includes only the June admission. The July admissions are excluded to avoid assigning a single death to two 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 2b5.

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

☐ **No risk adjustment or stratification**

☒ **Statistical risk model with 24 risk factors**

☐ **Stratification by risk categories**

☐ **Other,**

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

See risk model specifications in Section 2b3.4a and the attached data dictionary.

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

N/A. This measure is risk adjusted.

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

Selecting Risk Variables

Our goal in selecting risk factors for adjustment was to develop parsimonious models that included clinically relevant variables strongly associated with the risk of mortality in the 30 days following an index admission. We used a two stage approach, first identifying the comorbidity or clinical status risk factors that were most important in predicting the outcome, then considering the potential addition of social risk factors.

The original measure was developed with ICD-9. When ICD-10 became effective in 2015, we transitioned the measure to use ICD-10 codes as well. ICD-10 codes were identified using 2015 GEM mapping software. We then enlisted the help of clinicians with expertise in relevant areas to select and evaluate which ICD-10 codes

map to the ICD-9 codes used to define this measure during development. A code set is attached in field S.2b. (Data Dictionary).

For risk model development, we started with Condition Categories (CCs) which are part of CMS's Hierarchical Condition Categories (HCCs). The current HCC system groups the 70,000+ ICD-10-CM and 17,000+ ICD-9-CM codes into larger clinically coherent groups (201 CCs) that are used in models to predict mortality or other outcomes (Pope et al. 2001; 2011). The HCC system groups ICD- codes into larger groups that are used in models to predict medical care utilization, mortality, or other related measures.

To select candidate variables, a team of clinicians reviewed all CCs and excluded those that were not relevant to the Medicare population or that were not clinically relevant to the mortality outcome (for example, attention deficit disorder, female infertility). All potentially clinically relevant CCs were included as candidate variables and, consistent with CMS's other claims-based mortality measures, some of those CCs were then combined into clinically coherent CC groupings.

To inform final variable selection, a modified approach to stepwise logistic regression was performed. The Development Sample was used to create 1,000 "bootstrap" samples. For each sample, we ran a logistic stepwise regression that included the candidate variables. The results (not shown in this report) were summarized to show the percentage of times that each of the candidate variables was significantly associated with mortality ($p<0.01$) in each of the 1,000 repeated samples (for example, 90 percent would mean that the candidate variable was selected as significant at $p<0.01$ in 90 percent of the times). We also assessed the direction and magnitude of the regression coefficients.

The clinical team reviewed these results and decided to retain risk adjustment variables above a predetermined cutoff, because they demonstrated a strong and stable association with risk of mortality and were clinically relevant. Additionally, specific variables with clinical relevance to the risk of mortality were forced into the model (regardless of percent selection) to ensure appropriate risk adjustment for HF. These included variables representing markers for end of life/frailty, such as:

- Metastatic and other severe cancers (CC 8-CC 9)
- Hemiplegia, paraplegia, paralysis, functional disability (CC 70-CC 74, CC 103, CC 104, CC 189-CC 190)
- Stroke (CC 99-CC 100)
- Chronic kidney disease, stage 5 (CC 136)
- End-stage liver disease (CC 27)
- Hip fracture/dislocation (CC 170)

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

Social Risk Factors

We weigh SRF adjustment using a comprehensive approach that evaluates the following:

- Well-supported conceptual model for influence of SRFs on measure outcome (detailed below);
- Feasibility of testing meaningful SRFs in available data (section 1.8); and
- Empiric testing of SRFs (section 2b3.4b).

Below, we summarize the findings of the literature review and conceptual pathways by which social risk factors may influence risk of the outcome, as well as the statistical methods for SRF empiric testing. 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) and the Department of Health and Human Services Assistant Secretary for Policy and Evaluation (ASPE).

Causal Pathways for Social Risk Variable Selection

Although some recent literature evaluates the relationship between patient SRFs and the mortality outcome, few studies directly address causal pathways or examine the role of the hospital in these pathways (see, for example, Chang et al 2007; Gopaldas et al., 2009; Kim et al., 2007; LaPar et al., 2010; 2012; Lindenauer et al., 2013; Trivedi et al., 2014; Buntin et al., 2017; Lahewala et al., 2018; Kosar et al., 2020). Moreover, the current literature examines a wide range of conditions and risk variables with no clear consensus on which risk factors demonstrate the strongest relationship with mortality.

The social risk factors that have been examined in the literature can be categorized into three domains: (1) patient-level variables, (2) neighborhood/community-level variables, and (3) hospital-level variables.

Patient-level variables describe characteristics of individual patients, and include the patient's income or education level (Eapen et al., 2015). Neighborhood/community-level variables use information from sources such as the American Community Survey as either a proxy for individual patient-level data or to measure environmental factors. Studies using these variables use one dimensional measures such as median household income or composite measures such as the AHRQ-validated SES index score (Blum et al., 2014). Some of these variables may include the local availability of clinical providers (Herrin et al., 2015; Herrin et al., 2016). Hospital-level variables measure attributes of the hospital which may be related to patient risk (Roshanghalb et al., 2019). Examples of hospital-level variables used in studies are ZIP code characteristics aggregated to the hospital level or the proportion of Medicaid patients served in the hospital (Gilman et al., 2014; Jha et al., 2013).

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

1. **Patients with social risk factors may have worse health at the time of hospital 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 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. **Patients with social risk factors often receive care at lower quality hospitals**. Patients of lower income, lower education, or unstable housing have inequitable access to high quality facilities, in part 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. **Patients with social risk factors may receive 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. **Patients with social risk factors may experience worse health outcomes beyond the control of the health care system.** 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 financial priorities which don't allow for adequate recuperation or access to needed treatments, or a lack of access to care outside of the hospital.

Although we analytically aim to separate these pathways to the extent possible, we acknowledge that risk factors often act on multiple pathways, and as such, individual pathways are complex to distinguish analytically. Further, some social risk factors, despite having a strong conceptual relationship with worse

outcomes, may not have statistically meaningful effects on the risk model. They also have different implications on the decision to risk adjust or not.

Based on this model and the considerations outlined in section 1.8 – namely, that the AHRQ SES index and dual eligibility variables aim to capture the SRFs that are likely to influence these pathways (income, education, housing, and community factors) - the following social risk variables were considered for risk-adjustment:

- Dual eligible status
- AHRQ SES index

References

Barnato AE, Lucas FL, Staiger D, Wennberg DE, Chandra A. Hospital-level Racial Disparities in Acute Myocardial Infarction Treatment and Outcomes. Medical care. 2005;43(4):308-319.

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

Buntin MB, Ayanian JZ. Social Risk Factors and Equity in Medicare Payment. *New England Journal of Medicine.* 2017;376(6):507-510.

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. J Gen Intern Med. 2013 Feb; 28(2):269-82. doi: 10.1007/s11606-012-2235-x. Epub 2012 Oct 6.

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.

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

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.

Department of Health and Human Services, Office of the Assistant Secretary of Planning and Evaluation (ASPE). Second Report to Congress: Social Risk Factors and Performance in Medicare's Value-based Purchasing Programs. 2020; https://aspe.hhs.gov/system/files/pdf/263676/Social-Risk-in-Medicare%E2%80%99s-VBP-2nd-Report.pdf. Accessed July 2, 2020.

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.

Gilman M, Adams EK, Hockenberry JM, et al. California safety-net hospitals likely to be penalized by ACA value, readmission, and meaningful-use programs. Health Aff (Millwood). Aug 2014; 33(8):1314-22.

Gopaldas 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.

Hasnain-Wynia R, Kang R, Landrum MB, Vogeli C, Baker DW, Weissman JS. Racial and ethnic disparities within and between hospitals for inpatient quality of care: an examination of patient-level Hospital Quality Alliance measures. Journal of health care for the poor and underserved. May 2010;21(2):629-648.

Herrin J, Kenward K, Joshi MS, Audet AM, Hines SJ. Assessing Community Quality of Health Care. Health Serv Res. 2016 Feb;51(1):98-116. doi: 10.1111/1475-6773.12322. Epub 2015 Jun 11. PMID: 26096649; PMCID: PMC4722214.

Herrin J, St Andre J, Kenward K, Joshi MS, Audet AM, Hines SC. Community factors and hospital readmission rates. Health Serv Res. 2015 Feb;50(1):20-39. doi: 10.1111/1475-6773.12177. Epub 2014 Apr 9. PMID: 24712374; PMCID: PMC4319869.

Jha AK, Orav EJ, Epstein AM. Low-quality, high-cost hospitals, mainly in South, care for sharply higher shares of elderly black, Hispanic, and medicaid patients. Health affairs 2011; 30:1904-11.

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.

Kosar CM, Loomer L, Ferdows NB, Trivedi AN, Panagiotou OA, Rahman M. Assessment of Rural-Urban Differences in Postacute Care Utilization and Outcomes Among Older US Adults. *JAMA Netw Open*. 2020;3(1): e1918738. Published 2020 Jan 3. doi:10.1001/jamanetworkopen.2019.18738.

Krumholz HM, Brindis RG, Brush JE, et al. Standards for statistical models used for public reporting of health outcomes. Circulation. 2006; 113: 456-462. Available at: http://circ.ahajournals.org/content/113/3/456.full.pdf+html. Accessed January 14, 2020.

Lahewala S, Arora S, Tripathi B, et al. Heart failure: Same-hospital vs. different-hospital readmission outcomes [published correction appears in Int J Cardiol. 2020 Jun 15; 309:100]. *Int J Cardiol*. 2019; 278:186-191. doi: 10.1016/j.ijcard.2018.12.043.

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

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

Lindenauer PK, Lagu T, Rothberg MB, et al. Income inequality and 30-day outcomes after acute myocardial infarction, heart failure, and pneumonia: retrospective cohort study. BMJ. 2013 Feb 14; 346: f521. doi: 10.1136/bmj. f521.

Normand S-LT, Shahian DM. Statistical and Clinical Aspects of Hospital Outcomes Profiling. 2007/05 2007:206-226.

Pope GC, Ellis RP, Ash AS, et al. Diagnostic cost group hierarchical condition category models for Medicare risk adjustment. Final Report to the Health Care Financing Administration under Contract Number 500-95-048. 2000; http://www.cms.hhs.gov/Reports/downloads/pope_2000_2.pdf. Accessed February 25, 2020.

Pope GC, Kautter J, Ingber MJ, et al. Evaluation of the CMS-HCC Risk Adjustment Model: Final Report. 2011; https://www.cms.gov/Medicare/Health-Plans/MedicareAdvtgSpecRateStats/downloads/evaluation_risk_adj_model_2011.pdf. Accessed February 25, 2020.

Reames BN, Birkmeyer NJ, Dimick JB, et al. Socioeconomic disparities in mortality after cancer surgery: failure to rescue. JAMA surgery 2014; 149:475-81.

Roshanghalb A, Mazzali C, Lettieri E. Multi-level models for heart failure patients' 30-day mortality and readmission rates: the relation between patient and hospital factors in administrative data. *BMC Health Serv Res*. 2019;19(1):1012. Published 2019 Dec 30. doi:10.1186/s12913-019-4818-2.

Skinner J, Chandra A, Staiger D, et al. Mortality after acute myocardial infarction in hospitals that disproportionately treat black patients. Circulation 2005; 112:2634-41.

Trivedi AN, Nsa W, Hausmann LR, et al. Quality and equity of care in U.S. hospitals. The New England journal of medicine 2014; 371:2298-308.

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?

The table below shows the final variables in the model in the testing dataset with associated odds ratios (OR) and 95 percent confidence intervals (CI).

Table 4. Adjusted OR and 95% CIs for the HF Mortality Hierarchical Logistic Regression Model over Different Time Periods in the **Testing Dataset**

Variable	07/2016-06/2017 OR (95% CI)	07/2017-06/2018 OR (95% CI)	07/2018-06/2019 OR (95% CI)	07/2016-06/2019 OR (95% CI)
Age minus 65 (years above 65, continuous)	1.05 (1.05-1.05)	1.05 (1.05-1.05)	1.05 (1.05-1.05)	1.05 (1.05-1.05)
Male	1.18 (1.16-1.21)	1.21 (1.19-1.24)	1.23 (1.20-1.26)	1.22 (1.20-1.23)
History of coronary artery bypass graft (CABG) surgery	1.06 (1.03-1.09)	1.08 (1.05-1.11)	1.08 (1.05-1.11)	1.07 (1.05-1.09)
History of percutaneous transluminal coronary angioplasty (PTCA)	0.87 (0.84-0.89)	0.90 (0.87-0.92)	0.90 (0.87-0.92)	0.88 (0.87-0.90)
Metastatic cancer, acute leukemia, and other severe cancers (CC 8-9)	1.72 (1.66-1.79)	1.73 (1.66-1.80)	1.69 (1.62-1.75)	1.72 (1.68-1.76)
Diabetes mellitus (DM) or DM complications except proliferative retinopathy (CC 17-19, 123)	0.97 (0.95-0.99)	0.96 (0.94-0.98)	0.97 (0.95-0.99)	0.97 (0.96-0.98)
Protein-calorie malnutrition (CC 21)	1.98 (1.93-2.04)	2.00 (1.95-2.06)	2.00 (1.95-2.06)	2.02 (1.99-2.06)
Chronic liver disease (CC 27-29)	1.55 (1.47-1.62)	1.55 (1.48-1.62)	1.51 (1.44-1.58)	1.55 (1.51-1.59)
Dementia or other specified brain disorders (CC 51-53)	1.39 (1.35-1.42)	1.38 (1.35-1.41)	1.43 (1.40-1.47)	1.40 (1.38-1.42)
Major psychiatric disorders (CC 57-59)	1.01 (0.98-1.05)	1.02 (0.98-1.05)	0.95 (0.92-0.99)	1.00 (0.98-1.02)
Hemiplegia, paraplegia, paralysis, functional disability (CC 70-74, 103-104, 189-190)	1.11 (1.06-1.15)	1.09 (1.05-1.14)	1.09 (1.05-1.14)	1.10 (1.07-1.13)
Cardio-respiratory failure and shock (CC 84 plus ICD-10-CM codes R09.01 and R09.02, for discharges on or after October 1, 2015; CC 84 plus ICD-9-CM diagnosis codes 799.01 and 799.02, for discharges prior to October 1, 2015)	1.22 (1.19-1.25)	1.23 (1.20-1.26)	1.26 (1.23-1.30)	1.23 (1.21-1.25)

Variable	07/2016-06/2017 OR (95% CI)	07/2017-06/2018 OR (95% CI)	07/2018-06/2019 OR (95% CI)	07/2016-06/2019 OR (95% CI)
Congestive heart failure (CC 85)	1.18 (1.15-1.22)	1.19 (1.15-1.22)	1.16 (1.13-1.20)	1.18 (1.16-1.20)
Acute myocardial infarction (CC 86)	1.21 (1.17-1.25)	1.18 (1.15-1.22)	1.20 (1.16-1.24)	1.20 (1.17-1.22)
Unstable angina and other acute ischemic heart disease (CC 87)	0.95 (0.92-0.99)	0.98 (0.94-1.02)	0.98 (0.95-1.02)	0.98 (0.96-1.00)
Coronary atherosclerosis or angina (CC 88-89)	0.99 (0.96-1.01)	0.94 (0.91-0.96)	0.97 (0.94-0.99)	0.97 (0.95-0.98)
Valvular and rheumatic heart disease (CC 91)	1.09 (1.07-1.12)	1.10 (1.08-1.13)	1.10 (1.07-1.12)	1.10 (1.09-1.12)
Hypertension (CC 95)	0.74 (0.71-0.76)	0.75 (0.73-0.77)	0.74 (0.72-0.76)	0.75 (0.74-0.76)
Stroke (CC 99-100)	0.95 (0.91-0.99)	0.96 (0.92-0.99)	0.95 (0.91-0.99)	0.95 (0.93-0.97)
Vascular disease and complications (CC 106-108)	1.05 (1.03-1.08)	1.04 (1.01-1.06)	1.06 (1.04-1.08)	1.05 (1.04-1.07)
Chronic obstructive pulmonary disease (COPD) (CC 111)	1.11 (1.09-1.14)	1.08 (1.06-1.11)	1.07 (1.04-1.09)	1.09 (1.07-1.10)
Pneumonia (CC 114-116)	1.28 (1.25-1.31)	1.22 (1.19-1.24)	1.22 (1.20-1.25)	1.24 (1.22-1.26)
Renal failure (CC 135-140)	1.34 (1.31-1.38)	1.40 (1.37-1.44)	1.36 (1.32-1.39)	1.37 (1.35-1.39)
Trauma; other injuries (CC 166-168, 170-174)	1.11 (1.09-1.14)	1.10 (1.07-1.12)	1.12 (1.10-1.15)	1.11 (1.09-1.12)

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.**

Throughout this section, we present new SRF testing results based on the current testing dataset (2020); in addition, we show prior analyses included in the 2016 endorsement maintenance forms for comparison purposes.

SRFs	2020 Prevalence % (IQR)	2016 Prevalence % (IQR)
Dual	15.7% (9.60-25.0%)	13.2% (8.2-20.8%)
AHRQ Low SES	18.7% (6.50-35.7%)	17.6% (4.0-49.7%)

The prevalence of social risk factors in the HF cohort varies widely across measured entities in 2020. The median percentage of dual eligible patients was 15.7% (IQR 9.60%-25.0%) and the median percentage of patients with an AHRQ SES index score adjusted for cost of living at the census block group level equal to or below 42.7 (lowest quartile) was 18.7% (IQR 6.50%-35.7%) in 2020. These results are relatively consistent with the 2016 results presented above, though the prevalence of both SRFs has increased since 2016. The increase in dually eligible patients may be due to a refinement in the definition that occurred since 2016.

Comparison of observed mortality rates in patient with and without social risk in 2020 and 2016 (Table 6)

SRFs	2020 Observed Rate	2016 Observed Rate
Dual (vs. Non-Dual)	11.2% (vs. 11.4%)	10.8% (vs. 11.8%)
AHRQ Low SES (vs. SES score above 42.7)	10.4% (vs 11.8%)	10.9% (vs. 12.0%)

The patient-level observed HF mortality rates are lower for dual-eligible patients (11.2%) compared with 11.4% for non-dual patients in 2020. Similarly, the mortality rate for patients with an AHRQ SES index score equal to or below 42.7 are 10.4% compared with 11.8% for patients with an AHRQ SES index score above 42.7 in 2020. Patient-level mortality rates have declined among AHRQ low SES patients but not among dual-eligible patients.

Incremental effect of SRF variables in a multivariable model in 2020 and 2016

We examined the strength and significance of the SRF variables in the context of a multivariable model. When we include these variables in a multivariable model that includes all the claims-based clinical variables, the effect size of each of these variables is small. In 2020, dual eligibility and the AHRQ SES index have effect sizes (odds ratios) of 0.95 and 0.95 when added independently to the model, similar to 2016 findings. Furthermore, the effect size of each variable is slightly attenuated (0.94 and 0.96 for dual and AHRQ SES) when both are added to the model. Overall, the addition of these SRF variables to the model show that dually eligible and low SES patients have a lower risk of mortality when adjusted for other clinical variables.

We also find that the c-statistic is essentially unchanged with the addition of any of these variables into the model (Table 7), which is consistent with 2016 results.

Table 7

HF Mortality Models	2020 C-Statistic
Base Model: risk-adjusted model using the original clinical risk variables selected for the 2020 CMS public report of the HF mortality measure	0.69
Base Model plus AHRQ Low SES based on beneficiary residential 9-digit ZIP codes (SES9) as a social risk variable	0.69
Base Model plus dual eligibility (dual) as a social risk variable	0.69
Base Model plus SES9 and dual as social risk variables	0.69

Furthermore, we find that the addition of any of these variables into the hierarchical model has little to no effect on hospital performance. We examined the change in hospitals' RSMRs with the addition of any of these variables. The median absolute change in hospitals' RSMRs when adding a dual eligibility indicator is 0.007% (interquartile range [IQR] -0.010% – 0.005%) with a correlation coefficient between RSMRs for each hospital with and without dual eligibility added of 1.000. The median absolute change in hospitals' RSMRs when adding a low AHRQ SES Index score indicator to the model is 0.129% (IQR -0.096% – 0.148%) with a correlation

coefficient between RSMRs for each hospital with and without an indicator for a low AHRQ SES Index score is 0.981.

Summary

We also find that the impact of any of these indicators is small to negligible on model performance and hospital-level results. Given the controversial nature of incorporating such variables into a risk-model, we do not support doing so in a case that is unlikely to affect hospital profiling. Given these empiric findings, ASPE's recommendation to not risk adjust publicly reported quality measures for SRFs, and complex pathways which could explain the relationship between SRFs and mortality (and do not all support risk-adjustment), CMS chose to not incorporate SRF variables in this measure.

References:

Department of Health and Human Services, Office of the Assistant Secretary of Planning and Evaluation (ASPE). Second Report to Congress: Social Risk Factors and Performance in Medicare's Value-based Purchasing Programs. 2020; https://aspe.hhs.gov/system/files/pdf/263676/Social-Risk-in-Medicare%E2%80%99s-VBP-2nd-Report.pdf. Accessed July 2, 2020.

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

Approach to assessing model performance

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

Discrimination Statistics

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

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

Calibration Statistics

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

We tested the performance of the model for **the development dataset** 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*)**:**

Development and Validation Dataset:

1st half of randomly split development sample:

- C-statistic = 0.71
- Predictive ability (lowest decile %, highest decile %) = (3.0, 28.5)

2nd half of randomly split development sample:

- C-statistic = 0.70
- Predictive ability (lowest decile %, highest decile %) = (2.8, 29.0)

Results for the Testing Dataset

- C-statistic = 0.69
- Predictive ability (lowest decile %, highest decile %): (2.9, 25.1)

For comparison of model with and without inclusion of social risk factors, see above section.

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

For the development cohort, the results are summarized below:

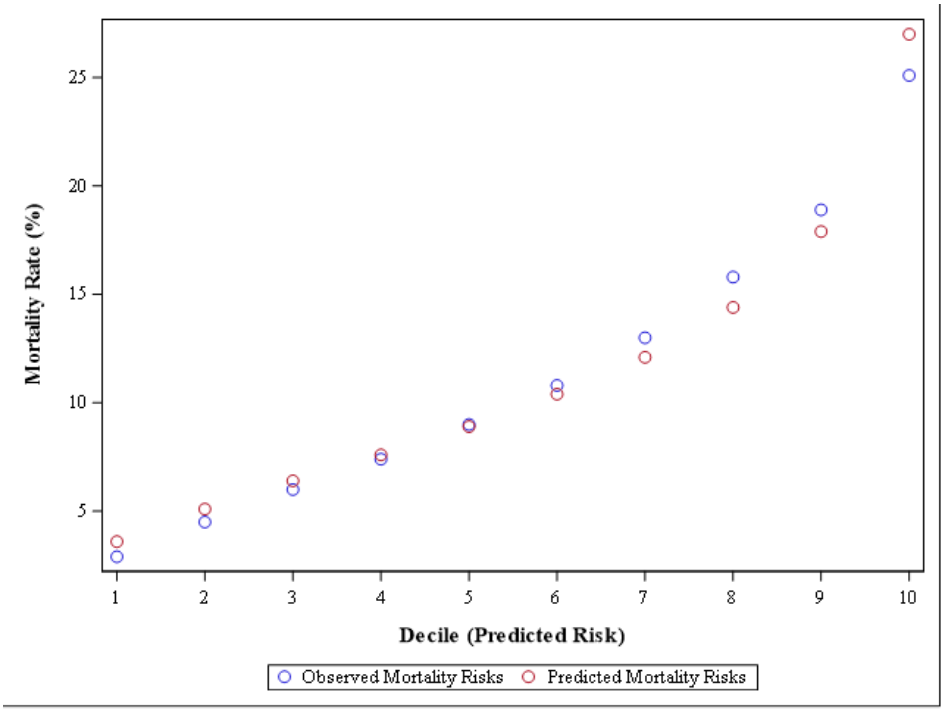
Development sample: Calibration: (0.0000, 1.0000)

Validation sample: Calibration: (-0.0035, 0.9928)

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

The risk decile plot (Figure 3) is a graphical depiction of the deciles calculated to measure predictive ability. Below, we present the risk decile plot showing the distributions for Medicare FFS data from July 2016 – June 2019 (Testing Dataset).

Figure 3. Risk Decile Plot



2b3.9. Results of Risk Stratification Analysis:

N/A

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

Discrimination Statistics

The C-statistic of 0.69 indicate moderate model discrimination. The model indicated a wide range between the lowest decile and highest decile, indicating the ability to distinguish high-risk subjects from low-risk subjects.

Calibration Statistics

Over-fitting (Calibration γ0, γ1)

If the γ0 in the validation samples are substantially far from zero and the γ1 is substantially far from one, there is potential evidence of over-fitting. The calibration value of close to 0 at one end and close to 1 to the other end indicates calibration of the model.

Risk Decile Plots

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

Overall Interpretation

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

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

NA

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

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

The measure score is hospital-specific risk-standardized mortality rates. These rates are obtained as the ratio of predicted to expected mortality, multiplied by the national unadjusted rate. The "predicted" mortality (the numerator) is calculated using the coefficients estimated by regressing the risk factors and the hospital-specific intercept on the risk of mortality. The estimated hospital-specific intercept is added to the sum of the estimated regression coefficients multiplied by the patient characteristics. The results are then transformed and summed over all patients attributed to a hospital to get a predicted value. The "expected" mortality (the denominator) is obtained in the same manner, but a common intercept using all hospitals in our sample is added in place of the hospital-specific intercept. The results are then transformed and summed over all patients in the hospital to get an expected value. To assess hospital performance for each reporting period, we re-estimated the model coefficients using the years of data in that period.

We characterize the degree of variability by:

1) Reporting the distribution of RSMRs.
    a. For public reporting of the measure, CMS characterizes the uncertainty associated with the RSMR by estimating the 95% interval estimate. This is similar to a 95% confidence interval but is calculated differently. If the RSMR's interval estimate does not include the national observed mortality rate (because it is lower or higher than the rate), then CMS is confident that the hospital's RSMR is different from the national rate and describes the hospital on the Hospital

Compare website as "better than the U.S. national rate" or "worse than the U.S. national rate." If the interval includes the national rate, then CMS describes the hospital's RSMR as "no different than the U.S. national rate" or "the difference is uncertain." CMS does not classify performance for hospitals that have fewer than 25 cases in the three-year period.

2) Providing the median odds ratio (MOR) (Merlo et al, 2006)
   a. The median odds ratio represents the median increase in the odds of mortality within 30 days of a HF admission date on a single patient if the admission occurred at a higher risk hospital compared to a lower risk hospital. MOR quantifies the between hospital variance in terms of odds ratio, it is comparable to the fixed effects odds ratio.

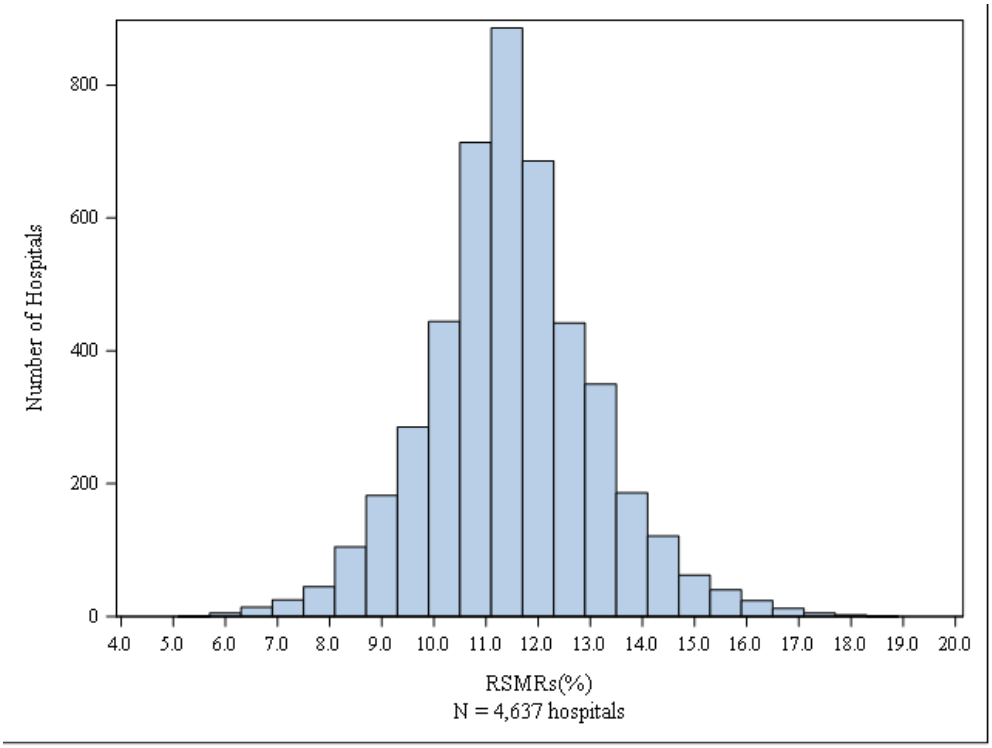
Reference

Merlo J, Chaix B, Ohlsson H, Beckman A, Johnell K, Hjerpe P, Råstam L, Larsen K. (2006) A brief conceptual tutorial of multilevel analysis in social epidemiology: Using measures of clustering in multilevel logistic regression to investigate contextual phenomena. J Epidemiol Community Health, 60(4):290-7.

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

Analyses of Medicare FFS data show substantial variation in RSMRs among hospitals.

Figure 4. Distribution (Histogram) Of Hospital-Level HF RSMRs



Out of 4,637 hospitals in the measure cohort, 259 performed "better than the U.S. national rate," 3,230 performed "no different from the U.S. national rate," and 156 performed "worse than the U.S. national rate" and 992 were classified as "number of cases too small" (fewer than 25) to reliably tell how well the hospital is performing.

The median odds ratio was 1.28.

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 median odds ratio suggests a meaningful increase in the risk of mortality if a patient is admitted with HF at a higher risk hospital compared to a lower risk hospital. A value of 1.28 indicates that a patient has a 28% increase in the odds of mortality at higher risk performance hospital compared to a lower risk hospital, indicating the impact of quality on the outcome rate is substantial.

The variation in rates and number of performance outliers suggests there remain differences in the quality of care received across hospitals for HF. This evidence supports continued measurement to reduce the variation.

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

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

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

factors in the risk adjustment model. However, if comparability is not demonstrated for measures with more than one set of specifications/instructions, the different specifications (e.g., for medical records vs. claims) should be submitted as separate measures.

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

N/A

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

N/A

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

N/A

2b6. MISSING DATA ANALYSIS AND MINIMIZING BIAS

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

The HF mortality measure used claims-based data for development and testing. There was no missing data in the development and testing data.

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

N/A

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

3. Feasibility

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

3a. Byproduct of Care Processes

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

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

Coded by someone other than person obtaining original information (e.g., DRG, ICD-9 codes on claims)

If other:

3b. Electronic Sources

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

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

ALL data elements are in defined fields in electronic claims

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

N/A

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

Attachment:

3c. Data Collection Strategy

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

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

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

This measure uses administrative claims data and enrollment data and as such, offers no data collection burden to hospitals or providers.

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)*.

N/A

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 Hospital Compare https://www.medicare.gov/hospitalcompare/search.html? Hospital Compare https://www.medicare.gov/hospitalcompare/search.html? Payment Program Hospital Value Based Purchasing (HVBP) Program https://www.qualitynet.org/inpatient/hvbp Hospital Value Based Purchasing (HVBP) Program https://www.qualitynet.org/inpatient/hvbp

*cell intentionally left blank

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

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

Public Reporting

Program Name, Sponsor: Hospital Compare, Centers for Medicare and Medicaid Services (CMS)

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

Payment Program

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

Purpose: The Hospital Value-Based Purchasing (HVBP) Program is a CMS initiative that rewards acute-care hospitals with incentive payments for the quality of care they provide to people with Medicare. It was established by the Affordable Care Act of 2010 (ACA), which added Section 1886(o) to the Social Security Act. The law requires the Secretary of the Department of Health and Human Services (HHS) to establish a value-based purchasing program for inpatient hospitals. To improve quality, the ACA builds on earlier legislation—the 2003 Medicare Prescription Drug, Improvement, and Modernization Act and the 2005 Deficit Reduction Act. These earlier laws established a way for Medicare to pay hospitals for reporting on quality measures, a necessary step in the process of paying for quality rather than quantity.

Geographic area and number and percentage of accountable entities and patients included: More than 3,000 hospitals across the country are eligible to participate in Hospital VBP. The program applies to subsection (d) hospitals located in the 50 states and the District of Columbia and acute-care hospitals in Maryland. More details about the Hospital VBP program are online at https://www.qualitynet.org/inpatient/hvbp.
The following hospitals are excluded from Hospital VBP:

- Hospitals and hospital units excluded from the Inpatient Perspective Payment System, such as psychiatric, rehabilitation, long-term care, children's, and cancer hospitals.
- Hospitals that are located in the state of Maryland participating in the Maryland All-Payer Model.
- Hospitals subject to payment reductions under the Hospital Inpatient Quality Reporting (IQR) Program.
- Hospitals cited by the Secretary of HHS for deficiencies during the performance period that pose an immediate jeopardy to patients' health or safety.
- Hospitals with an approved extraordinary circumstance exception specific to Hospital VBP; and
- Hospitals that do not meet the minimum number of cases, measures, or surveys required by Hospital VBP.

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

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

N/A, this measure is currently publicly reported

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

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

The exact number of measured entities (acute care hospitals) varies with each new measurement period. For the period between 2016 - 2019, all non-federal short-term acute care hospitals (including Indian Health Service hospitals), critical access hospitals, and VA hospitals (4,637 hospitals) were included in the measure calculation. Only those hospitals with at least 25 HF admissions were included in public reporting.

Each hospital generally receives their measure results in the Spring of each calendar year through CMS's QualityNet website. The results are then publicly reported on CMS's public reporting websites in the summer of each calendar year. Since the measure is risk standardized using data from all hospitals, hospitals cannot independently calculate their score.

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

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

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

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

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

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

During the summer of each year, the publicly reported measure results are posted on CMS's public reporting websites, a tool to find hospitals and compare their quality of care that CMS created in collaboration with organizations representing consumers, hospitals, doctors, employers, accrediting organizations, and other federal agencies. Measure results are updated in July of each calendar year.

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

Describe how feedback was obtained.

Question and Answer Inbox (Q&A)

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

Literature Reviews

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

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

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

For the HF mortality measure, we have received the following inquiries from hospitals since the last endorsement maintenance cycle:

1. Requests for detailed measure specifications including the ICD-9 and ICD-10 codes used to define the measure cohort or in the risk-adjustment model.
2. Requests for the SAS code used to calculate measure results.
3. Requests about the data source used to calculate the measure.
4. Questions about how transfers are handled in the measure calculation.
5. Requests for hospital-specific measure information such as HSRs; and
6. Requests for clarification of how inclusion and exclusion criteria are applied.

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

Summary of Question and Comments from Other Stakeholders:

For the HF mortality measure, we have received the following feedback from other stakeholders since the last endorsement maintenance cycle:

1. Requests for detailed measure specifications including the narrative specifications for the measure, CC-to-ICD-9 code crosswalks, and ICD-9 and ICD-10 codes used to define the measure cohort or in the risk-adjustment model.
2. Requests for the data source and the SAS code used to calculate measure results.
3. Requests for clarification of how inclusion and exclusion criteria are applied.
4. Queries about how cohorts and outcomes are defined, including how planned readmissions are defined.
5. Questions about how transfers are handled in the measure calculation; and
6. Requests for clarification on measure national rates.

Summary of Relevant Publications from the Literature Review:

Since the last endorsement cycle, we have reviewed more than 1,000 articles related to mortality following HF admissions. Relevant articles shared key themes related to spillover effects of the HF mortality measure on readmission rates for other conditions; considerations for additional risk adjustment variables, including social risk factors and other clinical comorbidities; association between public reporting of mortality rates and trends in mortality rates; potential unintended consequences of readmission measures on mortality outcomes; and the clinical differences between different types of HF.

Researchers have conducted considerable investigation of potential unintended consequences since the implementation of the Hospital Readmission Reductions Program. More specifically, the relationship between the implementation of the AMI, HF, and PN readmission measures in the Hospital Readmissions Reduction Program (HRRP) and subsequent trends in their respective mortality rates has been studied.

Some studies have argued that since HRRP implementation, readmissions for HF decreased but post-discharge mortality increased, suggesting a potential unintended consequence that readmission measures may be incentivizing hospitals to not readily admit patients with HF, and as a result, mortality rates increased (Gupta et al., 2018; Vaduganathan et al., 2019; Khera et al., 2018; Wadhera et al. 2018; Meyer et al., 2018). However, the same studies have acknowledged that HF mortality was increasing prior to HRRP implementation and that factors unrelated to HRRP could have caused this trend — for example, trends in hospice utilization, or the increasing use of do not resuscitate orders (DNRs), could lead to an increase in mortality rates. These findings

suggest that the increase in mortality (which, again, preceded HRRP) is not a result of denying admission to people seeking acute care services. Of note, other studies have found no apparent increase in HF mortality (Dharmarajan et al., 2017; MedPAC, 2018; Stensland., 2019).

Given the importance of this potential issue on patient outcomes, CMS commissioned an independent group to investigate whether there have been increases in mortality rates after HRRP implementation. CMS found through this investigation that no sufficient evidence exists to suggest that mortality has increased because of the HRRP readmission measures. CMS is committed to continuing to monitor trends in same-condition readmission and mortality rates through annual measure reevaluation and surveillance tasks.

References:

Dharmarajan K, Wang Y, Lin Z, et al. Association of Changing Hospital Readmission Rates With Mortality Rates After Hospital Discharge. JAMA. 2017;318(3):270-278.

Gupta A, Allen LA, Bhatt DL, et al. Association of the Hospital Readmissions Reduction Program Implementation With Readmission and Mortality Outcomes in Heart Failure. JAMA Cardiol. 2018;3(1):44-53.

Khera R, Dharmarajan K, Wang Y, et al. Association of the Hospital Readmissions Reduction Program With Mortality During and After Hospitalization for Acute Myocardial Infarction, Heart Failure, and Pneumonia. JAMA Netw Open. 2018;1(5): e182777.

Medicare Payment Advisory Commission. Mandated report: The effects of the Hospital Readmissions Reduction Program. Washington, DC 07/18 2018.

Meyer N, Harhay MO, Small DS, et al. Temporal Trends in Incidence, Sepsis-Related Mortality, and Hospital-Based Acute Care After Sepsis. Crit Care Med. 2018;46(3):354-360.

Stensland J. MedPAC evaluation of Medicare's Hospital Readmission Reduction Program: Update. In: 2019.

Wadhera RK, Joynt Maddox KE, Wasfy JH, Haneuse S, Shen C, Yeh RW. Association of the Hospital Readmissions Reduction Program With Mortality Among Medicare Beneficiaries Hospitalized for Heart Failure, Acute Myocardial Infarction, and Pneumonia. JAMA. 2018;320(24):2542-2552.

Vaduganathan M, McCarthy CP, Ayers C, et al. Longitudinal Trajectories of Hospital Performance across Targeted Cardiovascular Conditions in the United States. Eur Heart J Qual Care Clin Outcomes. 2019.

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

Each year, issues raised through the Q&A or in the literature related to this measure are considered by measure and clinical experts. Any issues that warrant additional analytic work due to potential changes in the measure specifications are addressed as a part of annual measure reevaluation. If small changes are indicated after additional analytic work is complete, those changes are usually incorporated into the measure in the next measurement period. If the changes are substantial, CMS may propose the changes through rulemaking and adopt the changes only after CMS received public comment on the changes and finalizes those changes in the IPPS or another rule. There were no questions or issues raised by stakeholders requiring additional analysis or changes to the measure since the last endorsement maintenance cycle.

Improvement

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

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

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

The median hospital 30-day, all-cause, RSMR for the HF mortality measure for the 3-year period between July 1, 2016 and June 30, 2019 was 11.4%. The median RSMR decreased by 0.7 absolute percentage points from July 2016-June 2017 (median RSMR: 11.6%) to July 2018-June 2019 (median: RSMR: 10.9%).

4b2. Unintended Consequences

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

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

We did not identify any unintended consequences during measure development or model testing. However, we are committed to monitoring this measure's use and assessing potential unintended consequences over time, such as the inappropriate shifting of care, increased patient morbidity and mortality, and other negative unintended consequences for patients.

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

N/A

5. Comparison to Related or Competing Measures

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

5. Relation to Other NQF-endorsed Measures

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

Yes

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

0330: Hospital 30-day, all-cause, risk-standardized readmission rate (RSRR) following heart failure (HF) hospitalization

0358: Heart Failure Mortality Rate (IQI 16)

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

1789: Hospital-Wide All-Cause Unplanned Readmission Measure (HWR)

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

3502: Hybrid Hospital-Wide (All-Condition, All-Procedure) Risk-Standardized Mortality Measure

3504: Claims-Only Hospital-Wide (All-Condition, All-Procedure) Risk-Standardized Mortality Measure

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

5a. Harmonization of Related Measures

The measure specifications are harmonized with related measures.

OR

The differences in specifications are justified

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

Are the measure specifications harmonized to the extent possible?

Yes

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

We did not include in our list of related measures any non-outcome (e.g., process) measures with the same target population as our measure. Our measure cohort was heavily vetted by clinical experts, a technical expert panel, and a public comment period. Additionally, the measure, with the specified cohort, has been publicly reported since 2008. Because this is an outcome measure, clinical coherence of the cohort takes precedence over alignment with related non-outcome measures. Furthermore, non-outcome measures are limited due to broader patient exclusions. This is because they typically only include a specific subset of patients who are eligible for that measure (for example, patients who receive a specific medication or undergo a specific procedure).

5b. Competing Measures

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

OR

Multiple measures are justified.

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

Describe why this measure is superior to competing measures (e.g., a more valid or efficient way to measure quality); OR provide a rationale for the additive value of endorsing an additional measure. (Provide analyses when possible.)

N/A

Appendix

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

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

Contact Information

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

Co.2 Point of Contact: Helen, Dollar-Maples, Helen.Dollar-Maples@cms.hhs.gov, 410-786-7214-

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

Co.4 Point of Contact: Doris, Peter, doris.peter@yale.edu

Additional Information

Ad.1 Workgroup/Expert Panel involved in measure development

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

The working group involved in the initial measure development is detailed in the original technical report available at www.qualitynet.org.

Our measure development team consisted of the following members:

Kanchana R. Bhat, M.P.H., Project Coordinator

Elizabeth E. Drye, M.D., S.M., Project Director

Harlan M. Krumholz, M.D., S.M., Principal Investigator

Sharon-Lise T. Normand, Ph.D., Co-Investigator*

Geoffrey C. Schreiner, B.S., Research Assistant

Yongfei Wang, M.S., Senior Statistical Analyst

Yun Wang, Ph.D., Senior Biostatistician

Measure Developer/Steward Updates and Ongoing Maintenance

Ad.2 Year the measure was first released: 2008

Ad.3 Month and Year of most recent revision: 07, 2019

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? 2020

Ad.6 Copyright statement: N/A

Ad.7 Disclaimers: N/A

Ad.8 Additional Information/Comments: N/A