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.

Purple text represents the responses from measure developers.

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

To navigate the links in the worksheet: Ctrl + click link to go to the link; ALT + LEFT ARROW to return

Brief Measure Information

NQF #: 3596

Corresponding Measures:

De.2. Measure Title: Hospital 30-day, all-cause, risk-standardized mortality rate (RSMR) following acute ischemic stroke hospitalization with claims-based risk adjustment for stroke severity

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

De.3. Brief Description of Measure: The measure estimates the hospital-level, risk-standardized mortality rate (RSMR) for patients discharged from the hospital with a principal discharge diagnosis of acute ischemic stroke. The outcome is all-cause 30-day mortality, defined as death from any cause within 30 days of the index admission date, including in-hospital death, for stroke patients. This is a re-specified measure with a cohort and outcome that is harmonized with the CMS's current publicly reported claims-based stroke mortality measure and includes the National Institutes of Health (NIH) Stroke Scale as an assessment of stroke severity upon admission in the risk-adjustment model. This measure uses Medicare fee-for-service (FFS) administrative claims for the cohort derivation, outcome, and risk adjustment.

1b.1. Developer Rationale: Stroke is the fifth most common cause of death, affecting approximately 795,000 people in the United States annually [CDC, 2020], and has a mortality rate of 17% [Go et al., 2014; Kochanek et al., 2014]. Stroke is also a leading cause of disability in the United States, which can lead to increased dependency on the health care system and higher subsequent costs associated with this care [Mozaffarian et al., 2015; CDC, 2020]. Mortality following stroke – an important adverse outcome that can be measured reliably and objectively, and that is influenced by the quality of care provided to patients during their initial hospitalization – is an appropriate measure of quality of care [DesHarnais et al., 1988; Weir et al, 2001]. Specifically, post-stroke mortality rates have been shown to be influenced by critical aspects of care such as response to complications, speediness of delivery of care, organization of care, and appropriate imaging [Hacke et al., 2004; Smith et al., 2006; Fang et al., 2008; Reeves et al., 2009; Lingsma et al., 2008; Hong et al., 2008; Fonarow et al., 2014; Bekelis et al., 2016; Xian et al., 2019; Jahan et al., 2019]. This work demonstrates the relationship between hospital organizational factors and performance on the stroke mortality measure and supports the ability of hospitals to impact these rates.

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 acute ischemic stroke. 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 outcome 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.

Rationale for Development of an Updated Claims-Only Stroke Mortality Measure

Current outcome measures use administrative claims data from the year prior to the index admission in the risk adjustment models. Stakeholders, including the AHA/ASA and other professional organizations, have highlighted the importance of including stroke severity in mortality measures for risk adjustment. Several studies have demonstrated that initial stroke severity is the strongest predictor of mortality in ischemic stroke patients [Smith et al., 2010; Nedeltchev et al., 2010; Fonarow et al., 2012].

This new claims-based stroke mortality measure addresses these stakeholder preferences and improves model performance by updating the current publicly reported claims-based stroke mortality measure to incorporate stroke severity scores into the risk-adjustment model. Advancements in clinical practice to incorporate new clinical assessments in administrative coding systems have made it possible to integrate these data into measures of hospital performance. The NIH Stroke Scale, which was created in 1989 and is widely used in routine stroke care, is collected in the GWTG-Stroke Registry, which has over 1,700 hospitals throughout the U.S. [Fonarow et al., 2014]. The NIH Stroke Scale is a 15-item neurologic examination stroke scale used to provide a quantitative assessment of stroke related neurologic deficit, by evaluating the effect of acute ischemic stroke on the levels of consciousness, language, neglect, visual-field loss, extraocular movement, motor strength, ataxia, dysarthria, and sensory loss. The NIH Stroke Scale is designed to be a simple, valid, and reliable tool that can be consistently administered at the bedside by physicians, nurses, or therapists. The use of the NIH Stroke Scale to assess stroke severity upon acute ischemic stroke patient presentation is recommended in the AHA/ASA Class I guidelines [Powers et al, 2019]. Furthermore, as of October 2016, the NIH Stroke Scale score is now coded as a secondary ICD-10-CM code within administrative claims, allowing it to be used in this measure. Inclusion of stroke severity data will not only address stakeholder preferences, but also improve the discrimination of the risk model.

References:

Adams HP, Jr., Davis PH, Leira EC, et al. Baseline NIH Stroke Scale score strongly predicts outcome after stroke: A report of the Trial of Org 10172 in Acute Stroke Treatment (TOAST).

Bekelis K, Marth NJ, Wong K, Zhou W, Birkmeyer JD, Skinner J. Primary Stroke Center Hospitalization for Elderly Patients with Stroke: Implications for Case Fatality and Travel Times. JAMA Intern Med. 2016;176(9):1361-1368.

Casper ML NI, Croft JB, Nilasena DS. Atlas of Stroke Hospitalizations Among Medicare Beneficiaries. 2008.

Centers for Disease Control and Prevention: Stroke. Available at: https://www.cdc.gov/stroke/index.htm. Accessed July 27, 2020.

DesHarnais SI, Chesney JD, Wroblewski RT, Fleming ST, McMahon LF, Jr. The Risk-Adjusted Mortality Index. A new measure of hospital performance. Med Care. Dec 1988;26(12):1129-1148.

Fang J, Keenan NL, Ayala C, Dai S, Merritt R, Denny CH. Awareness of stroke warning symptoms—13 states and the District of Columbia, 2005. MMWR. 2008;57(18):481–5.

Fonarow GC, Alberts MJ, Broderick JP, et al. Stroke outcomes measures must be appropriately risk adjusted to ensure quality care of patients: a presidential advisory from the American Heart Association/American Stroke Association. Stroke. May 2014;45(5):1589-1601.

Fonarow GC, Saver JL, Smith EE, et al. Relationship of national institutes of health stroke scale to 30-day mortality in Medicare beneficiaries with acute ischemic stroke. J Am Heart Assoc. Feb 2012;1(1):42-50.

Go AS, Mozaffarian D, Roger VL, et al. Heart disease and stroke statistics--2014 update: a report from the American Heart Association. Circulation. Jan 21, 2014;129(3): e28-e292.

Hacke W, Donnan G, Fieschi C, et al. Association of outcome with early stroke treatment: pooled analysis of ATLANTIS, ECASS, and NINDS rt-PA stroke trials. Lancet (London, England). Mar 6 2004;363(9411):768-774.

Hong KS, Kang DW, Koo JS, et al. Impact of neurological and medical complications on 3-month outcomes in acute ischaemic stroke. European journal of neurology: the official journal of the European Federation of Neurological Societies. Dec 2008;15(12):1324-1331.

Jahan R, Saver JL, Schwamm LH, et al. Association Between Time to Treatment with Endovascular Reperfusion Therapy and Outcomes in Patients with Acute Ischemic Stroke Treated in Clinical Practice. JAMA. 2019;322(3):252-263. doi:10.1001/jama.2019.8286.

Kochanek KD MS, Xu JQ, Arias E. Mortality in the United States, 2013. NCHS data brief, no. 178. 2014.

Lingsma HF, Dippel DW, Hoeks SE, et al. Variation between hospitals in patient outcome after stroke is only partly explained by differences in quality of care: results from the Netherlands Stroke Survey. Journal of neurology, neurosurgery, and psychiatry. Aug 2008;79(8):888-894.

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

Nedeltchev K, Renz N, Karameshev A, et al. Predictors of early mortality after acute ischaemic stroke. Swiss Medical Weekly. 2010;140(17-18):254-259.

Powers WJ, Rabinstein AA, Ackerson T, et al. Guidelines for the Early Management of Patients with Acute Ischemic Stroke: 2019 Update to the 2018 Guidelines for the Early Management of Acute Ischemic Stroke: A Guideline for Healthcare Professionals from the American Heart Association/American Stroke Association. Stroke. 2019;50(12): e344-e418.

Reeves MJ, Smith E, Fonarow G, Hernandez A, Pan W, Schwamm LH. Off-hour admission and in-hospital stroke case fatality in the get with the guidelines-stroke program. Stroke. Feb 2009;40(2):569-576.

Smith EE, Shobha N, Dai D, et al. Risk score for in-hospital ischemic stroke mortality derived and validated within the Get with the Guidelines-Stroke Program. Circulation. 2010;122(15):1496-1504.

Smith MA, Liou JI, Frytak JR, Finch MD. 30-day survival and rehospitalization for stroke patients according to physician specialty. Cerebrovascular diseases (Basel, Switzerland). 2006;22(1):21-26.

Weir NU, Sandercock PA, Lewis SC, Signorini DF, Warlow CP. Variations between countries in outcome after stroke in the International Stroke Trial (IST). Stroke. Jun 2001;32(6):1370-1377.

Xian Y, Xu H, O´Brien EC, et al. Clinical Effectiveness of Direct Oral Anticoagulants vs Warfarin in Older Patients with Atrial Fibrillation and Ischemic Stroke: Findings from the Patient-Centered Research into Outcomes Stroke Patients Prefer and Effectiveness Research (PROSPER) Study [published online ahead of print, 2019 Jul 22].

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 of the index admission for Medicare FFS patients aged 65 years and older with a principal discharge diagnosis of acute ischemic stroke.

S.6. Denominator Statement: The cohort includes inpatient admissions to all non-federal, short-term, acute care or critical access hospitals for Medicare FFS patients aged 65 years and older with a principal discharge diagnosis of acute ischemic stroke.

Additional details are provided in S.7 Denominator Details.

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

1. With inconsistent or unknown vital status or other unreliable data;
2. Enrolled in the Medicare hospice program at any time in the 12 months prior to the index admission, including the first day of the index admission; and

3. Discharged against medical advice (AMA).

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

De.1. Measure Type: Outcome

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

S.20. Level of Analysis: Facility

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

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

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

De.4. IF PAIRED/GROUPED, what is the reason this measure must be reported with other measures to appropriately interpret results? N/A

Preliminary Analysis: New Measure

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.

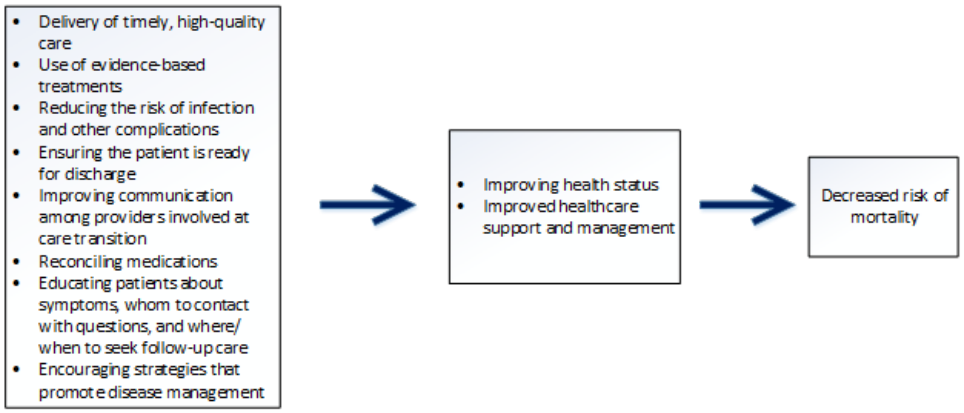
Criteria 1: Importance to Measure and Report

1a. Evidence

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

Evidence Summary

- **The developer provided the logic model below describing specific actions that can potentially be performed to impact stroke mortality:**
-



- The developer submitted that there is considerable literature linking post-stroke mortality rates to hospital organizational factors like the provider's response to complications, speediness of delivery of care, organization of care, coordinated transitions to the outpatient environment, antihypertensive and anticoagulant therapies, and appropriate imaging.
- Some specific examples provided by the developer:
    - Hospitals participating in quality improvement registries like Get With The Guidelines (GWTG) had lower in-hospital mortality rates among stroke patients than hospitals not participating in similar programs (Fonarow et al., 2014).
    - Patients being treated at hospitals participating in the GWTG quality improvement registry for stroke were significantly more likely to receive multiple evidence-based care interventions, such as tissue plasminogen activator (tPA) administration and evaluation by a neurologist (Howard et al., 2018).

Updates:

Question for the Committee:

- *Is there at least one thing that the provider can do to achieve a change in the measure results?*

Guidance from the Evidence Algorithm

Outcome measure (Box 1) -> Outcome and health action uses 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 used Medicare Fee for Service administrative claims data from October 2016 to June 2019 using hospitals where the NIH Stroke Scale was coded for 60% of the claims
    - **Sample Size**: 89,795 admissions at 329 hospitals.
    - **Mean risk standardized mortality rate (RSMR)** : 14.63%
    - **Median hospital RSMR**: 14.68%
    - **Range** :10.05% to 17.83%
    - **Interquartile range**: 13.82% to 15.52%.
- The developer included historical data from CMS's current publicly reported claims based stroke mortality measure in the IQR program from July 2013 to June 2016 and July 2016 to June 2019 to demonstrate a continued decline in mortality rates after the measure was implemented.
    - **Sample Size**: July 2013-June 2016: 519,732 admissions from 4,417 hospitals; July 2016-June 2019: 520,432 admissions from 4,254 hospitals.
    - **Median hospital RSMR**: July 2013-June 2016: 14.5%; July 2016-June 2019: 13.6%
    - **Interquartile range**: July 2013-June 2016: 14.0%-15.3%; July 2016-2019: 13.0% - 14.2%

Disparities

- The developer provides claims data using Medicare FFS claims, American Community Survey data, and Master Beneficiary Summary File (MBSF) data as data sources and categorized them by dual eligible status and AHRQ SES score. The data below was gathered between October 1, 2016 - June 30, 2019.

Measures	#hospitals	Percentage of patients	Minimum rate	Median rate	Maximum rate
Dual eligibles:Low proportion	83	<=7.37	11.72	14.53	17.21
Dual eligibles: High proportion	82	>=16.90	10.05	14.33	17.44
AHRQ SES score: Low proportion	81	<=5.22	12.34	14.52	16.16
AHRQ SES score: High proportion	82	>=23.12	11.76	14.51	17.83

Questions for the Committee:

- Is there a gap in care that warrants a national performance measure?
- Does the disparities information provided suggest that this measure should be risk-adjusted or stratified by SES?

Preliminary rating for opportunity for improvement: ☒ High ☐ Moderate ☐ Low ☐ Insufficient
Committee Pre-evaluation Comments:
Criteria 1: Importance to Measure and Report (including 1a, 1b, 1c)

1a. Evidence to Support Measure Focus: For all measures (structure, process, outcome, patient-reported structure/process), empirical data are required. How does the evidence relate to the specific structure, process, or outcome being measured? Does it apply directly or is it tangential? How does the structure, process, or outcome relate to desired outcomes? For maintenance measures–are you aware of any new studies/information that changes the evidence base for this measure that has not been cited in the submission? For measures derived from a patient report: Measures derived from a patient report must demonstrate that the target population values the measured outcome, process, or structure.

- use of the GWTG which reflects the NIHSS; there are challenges with the NIHSS but authors seem to have tested it when it is missing
- It seems that hospitals would be able to make changes such as improving time to needle, improving organization of care, coordinating transitions of care, giving appropriate medications to improve their performance on this measure.
- Not from a patient report.
- The measure would improve patient outcomes by providing patients, physicians, hospitals, and policy makers with information about hospital-level, risk-standardized mortality rates following hospitalization for acute ischemic stroke.
- The testing dataset included 329 non-federal, acute inpatient US hospitals that submitted claims with National Institute of Health (NIH) Stroke Scale scores for at least 60% of their stroke admissions.
- The dataset included administrative data on each patient for the 12 months prior to the index admission.

- 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.
- 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.
- This new claims-based stroke mortality measure addresses these stakeholder preferences and improves model performance by updating the current publicly reported claims-based stroke mortality measure to incorporate stroke severity scores into the risk-adjustment model.
- There are specific features that some hospitals may be able to provide (specialist care, endovascular therapy in a timely manner, but...my concern is can all hospitals realistically provide the care that has been shown to improve outcomes is this really under the hospitals control. the logic model identifies factors that the evidence doesn't support. I think this is the greatest weakness of this measure. However NQF has endorsed a similar measure looking at in hospital mortality so in the past we have not been concerned about that so...
- Related to the addition of baseline NIHSS, the measure directly applies, is meaningful to the target population.
- yes
- yes, developers focused on validity using data from the NIH Stroke Scale looking at Medicare claims and registry data and looked at validity assessed by external groups.
- Evidence is convincing
- There is sufficient evidence to support
- Quality of care seems very likely to directly apply to the measure
- RSMR following acute ischemic stroke can be a direct reflection of hospital care for stroke patients.

1b. Performance Gap: Was current performance data on the measure provided? How does it demonstrate a gap in care (variability or overall less than optimal performance) to warrant a national performance measure? Disparities: Was data on the measure by population subgroups provided? How does it demonstrate disparities in the care?

- interestingly, mortality is really not the issue but rather disability associated with stroke
- There is a range in results that suggests that improvement could be substantial. The disparities data are less clear.
- no
- Weaknesses:
- Unclear how many of the 329 hospitals already had Get With the Guidelines (GWTG) certification. (Patients being treated at hospitals participating in the GWTG quality improvement registry for stroke were significantly more likely to receive multiple evidence-based care interventions, such as tissue plasminogen activator (tPA) administration and evaluation by a neurologist).
- Is there a gap in care between GWTG Comprehensive and Primary certified hospitals? Measure does not specify between these classifications.
- Measure only looks at Medicare/65 and older patients. Consider adding younger stroke patients to improve outcome for that population. Note: GWTG does not include protocol for the pediatric population anywhere in the guidelines. The pediatric population is misdiagnosed 40% of the time at initial ED arrival. And delay in imaging ranges from 15 to 24 hours. https://www.ahajournals.org/doi/10.1161/STR.0000000000000183
- 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. It was not clear how the results of this data would be disseminated to patients/the community. Would emergency transport inform the patient of the hospital's stroke rating and would the patient/family be able to select the hospital to be transported to? How is this outcome measure different than the GWTG stroke rating of Comprehensive and Primary Stroke Centers? Strengths:

- The developer provides claims data using Medicare FFS claims, American Community Survey data, and Master Beneficiary Summary File (MBSF) data as data sources and categorized them by dual eligible status and AHRQ SES score.
- The inclusion of the NIH Stroke Scale has been shown to improve model discrimination for the publicly reported stroke mortality measure.
- Yes, the performance gap was identified and disparities was addressed and there appears to be disparities.
- The wide variation in measure performance for hospitals, and the evidence that this outcome is modifiable during the index hospitalization demonstrates an opportunity to reduce disparities in care
- yes
- the developers cite variability in mortality across age groups and suggest that this may be due in part to quality of care. They also relate this to readmissions and outcomes on readmission.
- Yes there is gap and there are disparities in the care justifying the need for this measure.
- What does it mean that the range for hospital RSMR went from 14.0-15.3% to 8.7-20.9% after the measure was introduced? Did introduction of the measure actually lead to worsening mortality rates at some hospitals?
- There seemed to be a relatively small range of predicted values for the measure; specifically the 10%-90% range was narrow, so it seems, depending on thresholds, not many hospitals may be singled out; I didn't see (but certainly could have missed) how many hospitals were rated as "better" or "worse" than national rate. There were no demonstrable SES disparities after adjustment for clinical variables, so SES adjustment was not included.
- Large performance gap was demonstrated by using historical claims data. Wide interquartile range suggests large opportunity for improvement. Whether the measure should be stratified by SES was less certain based upon the claims data provided.

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability: Specifications and Testing

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

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

Reliability

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

2a2. Reliability testing demonstrates if the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or

that the measure score is precise enough to distinguish differences in performance across providers. For maintenance measures – less emphasis if no new testing data provided.

Validity

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

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

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

Evaluators: NQF Scientific Methods Panel Subgroup

Methods Panel Review (Combined)

Methods Panel Evaluation Summary:

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

Reliability

- Reliability testing was conducted at the measure score level and the dataset used for testing included Medicare Parts A and B claims as well as the Medicare Enrollment Database (EDB).
- Signal to noise ratio testing was performed for both all hospitals and hospitals meeting the minimum case count of at least 25 cases for public reporting.
- The reliability score was 0.72; however had a wide range from 0.01 -- meaning that it was unreliable at some hospitals to 0.95 -- meaning that it was very reliable at others in the testing sample. The 25th and 75th percentiles were 0.51 and 0.83, respectively. Using the threshold of at least 25 cases which will be used for public reporting, the median reliability score was 0.75, yet still had a large range from 0.24 to 0.95 and 25th and 75th percentiles were 0.59 and 0.83, respectively.
- Specific comments from the Scientific Methods Panel on specifications and reliability are below, starting on Page 10.

Validity

- For data element validity, the developer compared the scores of the Medicare claims with the scores from GWTG-Stroke registry data and compared the scores using a sample size of 29,936 stroke hospitalizations. When comparing NIH Stroke scores to GWTG-Stroke Registry and administrative claims data, 93% was in 5 points of each other, and 84% of the data was within 2 points. The Pearson correlation coefficient between the 2 scores is 0.993 and weighed Kappa was 0.842.
- For construct validity, the developer also assessed the measure score correlation with the Overall Hospital Star Ratings Mortality measure Group score. The developers wanted to see if better performance on the measure was related to better performance on the Overall Star ratings for the hospitals. The overall correlation was 0.422.
- The measure developers used the National Quality Forum's guidelines, CMS Measure Management System (MMS) guidance, and AHA's "Standards for Statistical Models Used for Public Reporting of Health Outcomes" guideline to develop the measure.

- The social risk factors that were considered for this measure were AHRQ SES scores and dual eligible status, but the measure was risk adjusted for up to 20 age and clinical variables
- The Scientific Methods Panel comments on validity start at the bottom of page 12 below.

Questions for the Committee regarding reliability:

- 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:

- 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?

- limited; Kappa score for inter-rater reliability reported
- All seems appropriate
- still not convinced about reliability with admin database derivation
- • The data elements were clearly defined. The mortality measure exclusions were realistic. • The inclusion criteria were targeted and specific. • There are no concerns about the likelihood of this measure being consistently implemented.
- Reliability is hampered by availability of the NIH Stroke scale
- The NIHSS is increasingly being collected and reported, and given its use in the "new" measure, should encourage further adherence. Over all data elements reasonable (although I did not see cardiovascular disease and a few other potential modifiers of mortality in the data dictionary).
- co morbidities
- reliability seems to be well addressed.
- I am worried about the wide range of variability at different hospitals, but I am not sure exactly how to fix it…
- No concerns
- No evidence presented on the reliability of the chosen ICD-10 codes to identify strokes. If they had a merged dataset with GWTG, some sort of analysis could have been performed? If not reliably selected stroke patients, full stop; demonstrating such reliability seems foundational. I also could not understand which hospitals will be measured; just those with >60% NIHSS; those with >25 stroke discharges; who exactly. Finally, the great amount of missing NIHSS data is a concern; what is the mortality rate amongst the stroke cases without NIHSS scores coded vs. those with? If NIHSS is missing in a non-random fashion, then the model would be biased?
- It appears that the measure can reliably differentiate performance

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

- limited

- The variability in reliability scores is quite concerning.
- yes
- no
- If available no concerns for reliability
- No
- no
- no
- I am worried about the wide range of variability at different hospitals, but I am not sure exactly how to fix it…
- No concerns
- Only presented for NIHSS; none of the other ICD based diagnoses had reliability described.
- None

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

- looked at correlation correlations which was high between NIHSS and claims
- No
- yes
- no
- No
- No
- no
- no
- No concern
- No concerns
- Face validity refers to an advisory group; when did they last meet? I also recall a discussion about incomplete accounting for comfort measures; though they exclude patients on hospice prior to hospital admission, most stroke patients that die after transition to comfort measures only were NOT on hospice prior to their stroke; these patients ALL DIE, so any such case, which could in fact be very HIGH quality care, would count against a hospital - thus rejecting a claim to face validity.
- None

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 of 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?

- when looking at three social risk factors, it is unclear what community level means and if this relates to the hospital level
- I think this was done well.
- does not demonstrate outcomes relate to quality

- When comparing the mortality rates of the admissions with NIH Stroke Scale scores with all admissions after multiple imputation, the mortality rates were very similar, demonstrating that the multiple imputation approach is valid across the range of NIH Stroke Scale scores.
- The exclusions were appropriate for the over 65 age group and consistent with the evidence.
- To ensure the imputation of NIH Stroke Scale using multiple imputation was valid, they compared the distribution of the NIH Stroke Scale among the admissions with reported NIH Stroke Scale scores and the distribution of NIH Stroke Scale for all admissions after imputation, and compared their association with 30-day mortality.
- The social risk factors that were considered for this measure were AHRQ SES scores and dual eligible status, but the measure was risk adjusted for up to 20 age and clinical variables including income, education level, median household income and zip code.
- While dual-eligibility was associated with increased mortality in the bivariate analyses, the association was reversed after accounting for comorbidities and stroke severity, indicating that any effects of dual-eligibility are already covered by clinical risk variables.
- the risk adjustment has been requested by those being evaluated. It appears an appropriate step.
- As I noted, there may be other comorbidities which impact the measure (coexisting CAD, liver disease, COPD, etc.) but previous analyses may have shown that these do not significant impact the measure
- no
- ok
- No concern
- Appropriate
- Very well established approach by a strong statistical team, no doubt. SES variables did not add discrimination, so were not included in final model (which as a stroke provider i am proud of!). I've already mentioned the threats in other questions.
- The authors accounted for SES in the measurements development but found that it had little impact on the overall hospital scores. Social risk factors were thus ultimately excluded from the adjusted model.

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?

- differences about quality related to NIHSS and number of patients seen at a facility; similar results when using NIHSS and claims data
- Seems adequate
- I don't think it measures meaningful quality differences
- When comparing the NIH Stoke Scale scores within the GWTG-Stroke Registry and administrative claims data, 93% of the scores from the two data sources are within 5 points of each other and 84% are within 2 points.
- Stakeholders expressed concern that the currently implemented measure did not adjust for stroke severity, so the Yale New Haven Health Services Corporation/Center for Outcomes Research and Evaluation (YNHHSC/CORE) developed a stroke mortality measure which incorporated the NIH Stroke Scale to address these stakeholder concerns.

- There was substantial concordance between the NIH Stroke Scale scores within the GWTG-Stroke Registry and administrative claims data, as demonstrated by the close proximity of scores, as well as the Pearson 48 correlation coefficient of 0.993 and weighted kappa of 0.842
- For data element validity, the developer compared the scores of the Medicare claims with the scores from GWTG-Stroke registry data and compared the scores using a sample size of 29,936 stroke hospitalizations.
- When comparing NIH Stroke scores to GWTG-Stroke Registry and administrative claims data, 93% was in 5 points of each other, and 84% of the data was within 2 points.
- To evaluate whether performance of the risk measure is not overly biased by missing stroke severity data they analyzed the risk measure using multiple imputation for missing values.
- The testing sample was based on hospitals that report the NIH Stroke Scale for at least 60% of ischemic stroke admissions included 329 hospitals and 89,795 admissions. In this testing sample, NIH Stroke Scale scores were available in 71.71% of patients with an admission for ischemic stroke from October 1, 2016 to June 30, 2019.
- The provided material suggests that missing data does not make the measure invalid.
- Somewhat but well addressed in the preliminary analysis
- comorbidities as they relate to mortality risk
- no
- Again, my main concern is the variability of reliability.
- No
- Missing data is an issue, there are a lot of missing NIHSS scores, and no clear plan as to how to address other than assigning a zero if no present, thereby penalizing hospitals who do not report; unintended consequence there could be inaccurate, non-expert reporting of NIHSS scores, even gaming the system by reporting high scores. Are the patients with missing NIHSS the same as those with?
- If missing NIHSS is imputed as 0 there does appear to significant shift (about 24% of test population) to a neighboring quartile. How will missing NIHSS be treated in the ultimate measure?

Criterion 3. Feasibility

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

3. **Feasibility** is the extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement. The required data elements are in defined fields either from claims or from electronic sources.

   A noted limitation of the feasibility of this measure is that the NIH stroke score data is not kept in a national database for all non-federal acute care hospitals. Therefore, the feasibility of this measure depends on hospital's measuring NIH stroke scores and including those data in the claims. Collecting NIHSS information is a class I recommendation from AHA/ASA. Based on all the acute care hospitals from Oct 2016 to June 2019, NIHSS were available in 37% of admissions for acute ischemic stroke. This increased from 13% in Oct 2016 to 55.6% in May 2019, demonstrating increased availability of these data.

Questions for the Committee:

- Are the required data elements routinely generated and used during care delivery?

Preliminary rating for feasibility: ☐ High ☒ Moderate ☐ Low☐ Insufficient Committee Pre-evaluation Comments:

Criteria 3: Feasibility

3. **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 or other electronic sources)? What are your concerns about how the data collection strategy can be put into operational use?**
    - NIHSS used in care delivery
    - The concern about including the NIHSS in datasets is a good one. However, it appears the inclusion of that score is increasing. The application of this measure may also drive inclusion of the score.
    - not sure about data accuracy using surrogate of NIHSS
    - Currently there is no national database that includes NIH Stroke Scale score data for stroke patients admitted to all non-federal acute care hospitals. Therefore, implementation of this measure depends on hospitals including patients' NIH Stroke Scale scores, for all patients admitted with acute ischemic stroke, in the claims they submit to Medicare using ICD-10 codes. Collection of the NIH Stroke Scale is now Class I recommended in the AHA/ASA guidelines for care of patients admitted with acute ischemic stroke. New ICD-10 codes for NIH Stroke Scale scores became available to hospitals to include in Medicare claims, which are routinely collected as part of the billing process, in October 2016.
    - Again, stroke scale being determined is the concern
    - I think the measure is feasible with the sources of data readily available
    - none
    - data elements should be generated during routine care.
    - It is possible to calculate NIHSS scores from the chart and also this will stimulate sites to prospectively record NIHSS scores, which is a good thing.
    - No concerns
    - again, missing values abound for NIHSS
    - NIHSS should and is routinely collected on most patients admitted for stroke. The collection of required data elements is feasible.

Criterion 4: Usability and Use

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

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

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

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

Current uses of the measure

Publicly reported? ☐ Yes ☒ No

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

OR

Planned use in an accountability program? ☒ Yes ☐ No

Accountability program details [Accountability program(s) – details]

- Although, the measure is currently not in use, the developer plans to use this measure to replace the currently reported Hospital 30-Day Mortality Following Acute Ischemic Stroke Hospitalization measure. The initial measure did not risk adjust for stroke severity, so the new measure was created to account for those factors.
- In addition, the developer hosted a workgroup of medical experts in 2016 to provide feedback on the measure before posting the measure specifications for public comment.

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 Measured hospitals and other stakeholders are able to submit data to an email, which is replied to by the developer of this measure or CMS. The feedback to date have included requests to add the NIHSS to the risk adjustment model as well as questions about measure specifications. The developer has also been monitoring the literature on this topic, and since 2016, 250 articles have been published. Relevant articles shared key themes related to: considerations for additional risk adjustment variables, including social risk factors and other clinical comorbidities; national trends in stroke mortality and geographic variation; racial disparities in stroke mortality rates; comparison of stroke mortality rates in primary stroke centers (PSCs) compared to non-PSCs; and, examination of NIH Stroke Scale validation and impact on stroke mortality measure model performance.

Additional Feedback: N/A

Questions for the Committee:

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

Preliminary rating for Use: ☒ Pass ☐ No Pass

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

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

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

Improvement results

- The developer compares the median hospital RSMR between the 1st dataset (2013-2016) and the 2nd dataset (2016-2019). The median hospital RSMR in the 2013-2016 dataset was 14.5%. This dataset

included 519,732 admissions from 4,417 hospitals. The median hospital RSMR in the 2016-2019 combined dataset was 13.6% based on 520,432 admissions from 4,254 hospitals.

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

Unexpected findings (positive or negative) during implementation

- The developer did not identify any unintended consequences or unexpected benefits in the measure.

Potential harms

- No potential harms were listed at this time.

Additional Feedback:

- No additional feedback was added at this time.

Questions for the Committee:

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

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

- It appears feedback has been incorporated and addressed
- It appears that replacement of the Hospital 30-day Mortality Following Acute Ischemic Stroke Hospitalization would be an improvement because of the ability to adjust for stroke severity with the current measure.
- not sure
- Although, the measure is currently not in use, CMS plans to use this measure to replace the currently reported Hospital 30-Day Mortality Following Acute Ischemic Stroke Hospitalization measure. The initial measure did not risk adjust for stroke severity, so the new measure was created to account for those factors.
- Feedback given: considerations for additional risk adjustment variables, including social risk factors and other clinical comorbidities; national trends in stroke mortality and geographic variation; racial disparities in stroke mortality rates; comparison of stroke mortality rates in primary stroke centers (PSCs) compared to non-PSCs; and, examination of NIH Stroke Scale validation and impact on stroke mortality measure model performance.
- Yes
- Yes, it will replace an already publicly reported measure.
- yes
- seems appropriate

- No concern
- No concerns
- I did not see a clear plan for implementation, who will be measured, will it be based on having a % of NIHSS scores, or the "0" score penalty plan, or some imputation plan? Despite their claims otherwise, seems a lot of hospitals have their quintiles reclassified if NIHSS values are missing. There is a good description of feedback having been elicited, some dates attached to this feedback would have been better.
- Data is not currently publicly available but is planned to be used in an accountability program. Feedback was curated from all relevant stakeholders.

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.

- So much is dependent on the hospital itself which can lead to improvements based on data but limited due to variation in care despite certified stroke centers
- It appears that improvement is possible and no harms are anticipated.
- don't see how this is useful
- The developer compared the median hospital RSMR between the 1st dataset (2013-2016) and the 2nd dataset (2016-2019). The median hospital RSMR in the 2013-2016 dataset was 14.5%. This dataset included 519,732 admissions from 4,417 hospitals. The median hospital RSMR in the 2016-2019 combined dataset was 13.6% based on 520,432 admissions from 4,254 hospitals.
- 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. Unclear how the consumers are notified/informed as to what the performance is of any given hospital.
- No unintended consequences or unexpected benefits were identified.
- No identified unintended consequences beyond that of having inappropriate measures and the consequence of those.
- Similar to trauma centers, it will be important to monitor performance to ensure that severally ill patients that are transferred in an already debilitated state with little hope for intervention or improvement (or even survival), do not negatively impact the receiving center measure performance. Perhaps the variable "transfer from another hospital" will capture some of this.
- Beneficial to track
- The measure seems well thought in that benefits outweigh harms.
- Cannot think of any
- I wonder if the wider range for hospital RSMR is due to some unintended consequence of introducing the measure
- Unintended consequences could include post hoc inaccurate coding of NIHSS, also, hospital willingness to allow patients to progress to comfort measures only may be undermined if they knew their stroke mortality quality score would get dinged every time, and they may even lose some money doing it. That all being said, the measure might identify a set of high performing and low performing hospitals, the former perhaps being in a position to advise the latter in ways that could improve patient outcomes.

- Developers provided suggestive data that overtime the outcome has been improving with regular reporting of NIHSS measure. No unintended consequences were identified by the developer.

Criterion 5: Related and Competing Measures

Related or competing measures

- 0467 Acute Stroke Mortality Rate (IQI 17)
- 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
- Non NQF endorsed competing measure: Hospital 30-Day, All-Cause, Risk-Standardized Mortality Rate Following Acute Ischemic Stroke (not NQF endorsed)

Harmonization

Developer indicates that measure specifications and cohorts are harmonized to the furthest extent possible. Developer also expands on the complementary and related nature of NQF 3596 and 0467, but distinguishes them as non-competing; the developer acknowledges that both measures assess mortality for patients admitted to acute care hospitals with a principal discharge diagnosis of acute ischemic stroke, but notes the different focus on 30-day (NQF 3596) versus inpatient mortality (NQF 0467). They also state that the specified outcomes for each measure are different.

The developer noted that they did not include a list of related non-outcome measures with the same target population because of the importance of maintaining clinical coherence of the cohort across comparison.

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

5. 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?

- NA
- This appears to have been taken into consideration.
- unknown
- The competing Measure, Hospital 30-Day, All-Cause, Risk-Standardized Mortality Rate Following Acute Ischemic Stroke (not NQF endorsed) will be harmonized.
- This measure and the NQF endorsed Acute Stroke Mortality Rate (IQI 17) (AHRQ) Measure #0467 are complementary and related rather than competing measures.
- No concerns
- None that I am aware of
- No
- none that I am aware of
- Not to my knowledge
- No
- not clear to me how related measures are actually being used...
- The developers identified appropriate related or competing measures. There were no opportunities for further harmonization identified.

Public and Member Comments

Comments and Member Support/Non-Support Submitted as of: 01/15/2021

- Comment by: **American Medical Association**

  The American Medical Association (AMA) appreciates the opportunity to comment on #3596, Hospital 30-day, all-cause, risk-standardized mortality rate (RSMR) following acute ischemic stroke hospitalization with claims-based risk adjustment for stroke severity. We are disappointed to see the minimum measure score reliability results of 0.24 using a minimum case number of 25 patients. We believe that measures must meet minimum acceptable thresholds of 0.7 for reliability. We request that the Standing Committee evaluate whether the measure specifications with only a case minimum of 25 patients is acceptable and if the measure meets the reliability criterion

- Comment by: **Federation of American Hospitals**

  The Federation of American Hospitals (FAH) appreciates the opportunity to comment on Measure #3596, Hospital 30-day, all-cause, risk-standardized mortality rate (RSMR) following acute ischemic stroke hospitalization with claims-based risk adjustment for stroke severity. FAH is concerned that even though the median reliability score was 0.7 for hospitals with at least 25 cases, reliability ranged from 0.24 to 0.95 and believes that the developer must increase the minimum sample size to a higher number to produce a minimum reliability threshold of sufficient magnitude (e.g. 0.7 or higher). As a result, the FAH requests that the Standing Committee carefully consider whether the measure as specified meets the reliability criterion.

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

Combined Methods Panel Scientific Acceptability Evaluation

Scientific Acceptability: Preliminary Analysis Form

Measure Number: 3596
Measure Title: Hospital 30-day, all-cause, risk-standardized mortality rate (RSMR) following acute ischemic stroke hospitalization with claims-based risk adjustment for stroke severity
Type of measure:
☐ **Process** ☐ **Process: Appropriate Use** ☐ **Structure** ☐ **Efficiency** ☐ **Cost/Resource Use**
☒ **Outcome** ☐ **Outcome: PRO-PM** ☐ **Outcome: Intermediate Clinical Outcome** ☐ **Composite**
Data Source:
☒ **Claims** ☐ **Electronic Health Data** ☐ **Electronic Health Records** ☐ **Management Data**
☐ **Assessment Data** ☐ **Paper Medical Records** ☐ **Instrument-Based Data** ☒ **Registry Data**
☒ **Enrollment Data** ☒ **Other**
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: NONE

Panel Member #2: None

Panel Member #4: The measure specification requires hospital submission of the NIH Stroke Scale for at least 60% of stroke admissions. Even with a relatively low submission threshold only 329 hospitals met that criteria.

Panel Member #8: . Numerator is yes/no for any mortality. Denominator exclusions are standard for Medicare measures and clear.

Panel Member #9: No concerns

RELIABILITY: TESTING

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

3. **Reliability testing level ☒ Measure score ☒ Data element ☐ Neither**
4. **Reliability testing was conducted with the data source and level of analysis indicated for this measure ☒ Yes ☐ No**
5. If score-level and/or data element reliability testing was NOT conducted or if the methods used were NOT appropriate, was **empirical VALIDITY testing** of patient-level data conducted?

   ☒ Yes ☐ No
6. **Assess the method(s) used for reliability testing**

Submission document: Testing attachment, section 2a2.2

Panel Member #1: S/N analysis

Data element reliability examined by comparing administrative data and GWTG Stroke Registry

Panel Member #2: Signal-to-noise at the facility level was assessed.

Panel Member #4: The methods are somewhat under described.

Panel Member #5: There was clearly a good faith effort to establish reliability of the data elements and measure score.

Panel Member #6: Developers estimated the hospital-level reliability using signal-to-noise analysis. The variation between hospitals ('signal') comprises the total variation ('noise' and 'signal') in the outcome in this case because the reliability of any one hospital's measure score will vary depending on the number of patients. Hospitals with higher volume will tend to have more reliable scores, while those with lower volume will tend to have less reliable scores.

Panel Member #8: Measure score reliability testing was signal to noise at the facility level using the Adams formula.

Panel Member #9: Signal to noise reliability score for 329 hospitals. Measure Stewards marked that data element reliability was performed but I was unable to locate those results.

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

Submission document: Testing attachment, section 2a2.3

Panel Member #1: S/N: Median 0.75 Mean 0.70

S/N: 10th percentile 0.45

S/N: 90th percentile 0.88

GWTG Stroke registry and admin claims, 93% within 5 point and 84% within 2 points. Distributions overlap

The 10th-90th percentile RSMRs are 13.04-16.28, a 3.24 percentage point difference. Given the number of cases, this is clinically meaningful but the limited range imposes a potential burden to reliably measure relative performance.

The results of the testing suggests the measure can reliably differentiate performance. The Mean and median S/N reliability statistics equal or exceed the Adams S/N analysis 0.7 reliability defacto standard.) Somewhat concerned about reliability in low volume hospitals.

Panel Member #2: The median signal-to-noise reliability score was estimated to be 0.75, including these facilities that had qualifying discharge diagnosis of stroke for 25 patients or more.

Panel Member #4: Again the results are somewhat under described. There are measure score results reported for both all hospitals and hospitals with at least 25 cases. Which is the specification, zero or 25? The results are more compelling with a minimum case count of 25. There are no data element reliability results reported.

Panel Member #6: Signal-to-Noise: There were 89,795 admissions in the testing sample. Signal to noise reliability score was calculated for all hospitals in the testing sample (N=329) and hospitals with at least 25 cases (N=292) to isolate the hospitals with enough ischemic stroke admissions to receive a publicly reported RSMR. For all hospitals in the testing sample, the median reliability score was 0.72, ranging from 0.01 to 0.95. The 25th and 75th percentiles were 0.51 and 0.83, respectively. For hospitals with at least 25 cases, the median reliability score was 0.75, ranging from 0.24 to 0.95. The 25th and 75th percentiles were 0.59 and 0.83, respectively. The median reliability scores demonstrates moderate to good reliability.

Panel Member #8: For all 329 hospitals the median reliability was 0.72, with a range from 0.01 to 0.95, driven by volume. The 25th and 75th percentiles were 0.51 and 0.83. For those with at least 25 cases, the median reliability score was 0.75, ranging from 0.24 to 0.95. The 25th to 75th percentiles were 0.59 and 0.83.

Panel Member #9: Median reliability scores were stable between all hospitals and hospitals with >25 cases. Median score 0.72-0.75 sufficient reliability.

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)

Panel Member #6: I DO NOT SEE RELIABILITY TESTING RESULTS FOR DATA ELEMENTS, though submission indicates this level of reliability testing was completed

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

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

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

☐ **Low** (NOTE: Should rate LOW if you believe specifications are NOT precise, unambiguous, and complete or if testing methods/results are not adequate)

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

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

Panel Member #1: The results of the testing suggests the measure can reliably differentiate performance. The Mean and median S/N reliability statistics equal or exceed the Adams S/N analysis 0.7 reliability defacto standard.) Somewhat concerned about reliability in low volume hospitals.

Panel Member #2: The facility-level reliability of 0.75 based on signal-to-noise ratio analysis can be considered moderate.
Panel Member #4: The measure score reliability results are not reported stratified by hospital volume. Given the specification requirement for submission of the NIH stroke score one suspects only the largest hospitals reported. It is not possible given the material submitted to determine the reliability of the measure when reported on the universe of hospitals.
Panel Member #5: Reliability testing was adequate. The signal to noise value is adequate.
Panel Member #6: Results of signal-to-noise show high reliability. Not sure why they indicated data element reliability testing but did not report results? I don't think we can mark down because of that, however. There is some conflicting opinion in SMP as to whether if measure score is reliability, data elements are reliable as well.
Panel Member #7: "Measure Score Reliability Results
Using the approach used by Adams et. al., we obtained the median signal-to-noise reliability score of 0.75, which demonstrates sufficiently high reliability (Adams et al, 2010).
Reference:
Adams J, Mehrota, A, Thoman J, McGlynn, E. (2010). Physician cost profiling – reliability and risk of misclassification. NEJM, 362(11): 1014-1021."
Panel Member #8: Across all hospitals, the results are moderate, and also true for those with greater than 25 cases.
Panel Member #9: No concerns

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: Since the primary reliability measure is based on hospitals with a minimum of 25 cases, I wonder why the measure developer did not consider this as an exclusion restriction, and thus restrict the entire analyses to the 292 hospitals rather than 329 hospitals.
Panel Member #4: None
Panel Member #6: None
Panel Member #8: None.. Exclusions are about 1 % of all cases and consistent with other Medicare mortality measures.
Panel Member #9: No concerns

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: The 10th-90th percentile RSMRs are 13.04-16.28, a 3.24 percentage point difference. Given the number of cases, this is clinically meaningful but the narrow range imposes a potential burden to reliably measure relative performance.
Panel Member #2: I am still little confused how the measure developer is defining high-risk and low-risk hospitals when defining median odds ratio. I am guessing that they used the following method defined in section 2b4.1; however, I am not sure: *If the RSMR's interval estimate does not include the national observed mortality rate (is lower or higher than the rate), then CMS is confident that the hospital's RSMR is different from the national rate and describes the hospital on the Hospital Compare website as "better than the U.S. national rate" or "worse than the U.S. national rate."*
Panel Member #4: The variation in measures scores is very slight (less than 2%)
Panel Member #6: Medicare FFS data show some variation in RSMRs among hospitals based on the testing sample (October 1, 2016-June 30, 2019). The median hospital RSMR was 14.68%, with a range of 10.05% to 17.83% which is a bit tight but appears fairly normally distributed. The median odds ratio using the between hospital variance was 1.21. These results indicate the measure will be able to identify meaningful differences in performance of hospitals.
Panel Member #8: As stated in the submission, for within-hospital variance the methodology used employs an estimated hospital-specific intercept and then adds the sum of the estimated

regression coefficients multiplied by patient characteristics. This is then transformed and summed over all patients attributed to a hospital to get a predicted value. For the denominator, a common intercept across all hospitals is utilized. If the confidence interval includes the national rate, the conclusion is "no different" or "difference is uncertain. If the confidence interval does not include the national rate, then the conclusion is either "better than" or "worse than" the national rate.
For between hospital variance a median odds ratio is used.
Panel Member #9: No concerns

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 #2: I have significant concerns regarding replacing missing stroke severity scale with zeros. First, when the measure developer says that imputation of missing stroke severity scales with zeros leads to slight improvement of C-statistics to 0.76 from 0.75 (as currently reported in inpatient quality reporting by CMS), should this be a valid comparison? The C-statistics in the models in the testing environment used hospitals that reported NIH stroke scales for at least 60% of the patients. Shouldn't that be the threshold for comparison? Second, the developer also claims that *"most hospitals will remain in the same quintile or move to a neighboring quintile (92.8%) from the stroke mortality without risk adjustment for stroke severity currently reported within IQR".* However, of that number, approximately 24% of the hospitals will move to a neighboring quintile, and approximately 8% will move to a higher quintile of RSMR than reported in IQR. When this score is later going to be used for payments, these movements will have substantial financial consequence.
Panel Member #4: None
Panel Member #6: N/A
Panel Member #8: Not applicable
Panel Member #9: No concerns

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

Submission document: Testing attachment, section 2b6.
Panel Member #1: N/A
Panel Member #2: It would have been more insightful if the developer reported and compared the characteristics of hospitals that had at least 60% NIH severity scores vs those without, which would have informed us about any observed/measured systematic bias between these two types of hospitals.
Panel Member #4: None
Panel Member #6: None
Panel Member #8: Missing data is imputed based on other patient variables. There was little difference between mortality rates not imputed and those imputed
Panel Member #9: No concerns

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
Panel Member #5: ZIP code level—Area Deprivation Index (ADI) from Census data (2009-2013)
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

Panel Member #1: C-statistic of 0.86 for model, a solid risk adjustment score.

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: Risk adjustment approach is standard CMS HCC model. C-statistic indicated good performance of risk adjustment model. SRF's while differentiated across patients and to some extent across hospitals, only explain a small portion of variance and correlation of scores based on risk adjustment model with and without SRFs is ~0.999

Panel Member #2: The developers have done a good job explaining the risk-adjustment approach including the approach adopted with regard to social risks.

Panel Member #4: Social risk- factors are well conceptualized

Panel Member #5: Risk adjustment was generally adequate, though not exceptional. Presentation of "Social Risk Proportion" (section 1b.4) was confusing. The analyses and discussion of the analyses for social risk factors was extensive. I would caution the Developers that the c-statistic (although the reported values were impressive) is an overall measure rather than a specific indicator of how social risk adjustment affects the most extreme Providers (i.e., those with very few or very many patients with particular socio-demographic risk factors). An analysis of how the inclusion of socio-demographic risk factors affects the risk adjusted performance of these extreme Providers would be more meaningful than simple differences in c-statistics for the overall population.

Panel Member #6: The final patient-level risk-adjustment model included 20 variables. The candidate variables for the model were derived from: the index admission, with comorbidities identified from the index admission secondary diagnoses (excluding potential complications); 12-month pre-index inpatient data (for any condition); outpatient hospital data; and Part B physician data. Variables were selected using the logistic regression model with the stepwise selection method based on 1,000 bootstrapping samples. Risk variables were retained if they were significantly associated with mortality ($p<0.01$) in 90 percent of the 1,000 repeated samples.

Social risk factors tested included dual eligible status and low AHRQ SES index. They assessed the relationship between the 2 social risk factor variables and the outcome and examined the incremental effect in a multivariable model. They also examined the extent to which the addition of any one of these variables improved model performance or changed hospital results. The median percentage of dual-eligible patients is 11.19% (interquartile range [IQR] 7.69%-16.85%). The median percentage of patients with a low AHRQ SES index score (lowest quartile) is 12.03% (IQR 5.26%-23.08%). The patient-level observed stroke mortality rates are higher for dual-eligible patients (17.49%) compared with all other patients (14.14%). The mortality rate for patients with a low AHRQ SES index score was slightly higher (15.27%) compared with all other patients (14.42%). The dual-eligibility variable had an odds ratio of 1.29 (95% CI 1.22, 1.36) in the bivariate analysis and 0.92 (95% CI 0.86, 0.98) in the multivariate analysis (which is the opposite effect expected based on bivariate model indicating the effect is covered by other clinical risk variables). The low AHRQ SES variable had an odds ratio of 1.07 (95% CI 1.02, 1.12) in the bivariate analysis and 1.04 (95% CI 0.98, 1.11) and significant in the multivariate analysis.

In all cases the c-statistics for the stroke patient-level multivariate models with the SES variables in the models were unchanged from those without (model with original variables: 0.86; model with dual-eligible variable: 0.86; model with AHRQ SES index variable: 0.86).

They also found little change in hospital scores when including these social risk factors. The mean absolute change in hospitals' RSMRs when adding a dual-eligibility indicator was 0.001% with a correlation coefficient between RSMRs for each hospital with and without dual-eligibility of 0.999. The mean absolute change in hospitals' RSMRs when adding a low SES AHRQ indicator was 0.00% with a correlation coefficient between RSMRs for each hospital with and without low SES of 0.999.

The final decision was to not include the social risk factors in the risk adjustment model.

The model was evaluated using the C-statistic, which was 0.86 indicating strong predictive ability.

Panel Member #8: A team of experts reviewed all Condition Categories and selected the initial candidate variables. These were then subject to a logistic stepwise regression to determine association with the

model outcome. Then stroke severity was added to end up with 20 variables, including the NIH Stroke Scale Score. Social Risk factors were analyzed to assess contribution to the score, but did not significantly improve the model.

Panel Member #9: Thorough analysis, no concerns

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: Correlation with STAR ratings mortality score. Assessment of impact of using imputed data for Stroke Risk. Data element validity compared data from Mcare claims and GWTG Stroke Registry. Expert and stakeholder input, public comment.

Panel Member #2: Although not checked in the submission form for critical data element validation, the developer validated the NIH Stroke scale with the Get with the Guideline Stroke Registry.
Empirical validation of risk standardized mortality rate (RSMR), the outcome of this measure, was conducted by comparing CMS' Overall Star Ratings Mortality Measure Group Score.
Face validity was assessed through multiple ways: first, the measure was developed in consultation with national guidelines/guidance from NQF, CMS Measure Management System (MMS) guidance, and the guidance articulated in the American Heart Association scientific statement. Secondly, the measure developer obtained expert and stakeholder input via regular discussions with an advisory working group and a 30-day public comment period in order to increase transparency and to gain broader input into the measure.

Panel Member #4: The data element validity analysis is completing. The measure score validity testing is the lowest maturity level. The face validity is nonexistent. There seems to be a suggestion of "process validity" where if a measure is developed using a specific process then it is validly by default. That is not very compelling.

Panel Member #6: Data element reliability was assessed using the GWTG-Stroke Registry, comparing the NIH Storke Scale scores coded in claims to the registry score derived from patients' medical records. Empirical validity for the measure was further evaluated using correlation to the Overall Hospital Quality Star Ratings Mortality Group score which targets the same domain of quality. Validity was also assessed by external groups. Stakeholders had previously expressed concern that the current measure did not adjust for stroke severity.

Panel Member #8: Data validity was based on the NIH Stroke Scale to re-specify this measure. Then, a comparison was made between claims data and the data derived from the Stroke Registry. There was 93% score concordand within 5 points and 84% within 2 points.
Measure score validity was based on correlations with with the Star Ratings Mortality Measure Group Score and expert opinion. Correlation with the Star Ratings was 0.422.

Panel Member #9: Compared performance on the stroke mortality measure scores (RSMR) to performance quintiles on the Overall Star Ratings Mortality measure group score. In addition compared the NIH scores recorded in the claims data to the scores in the GWTG-Stroke Registry data to consider adding NIH scale as risk adjustment

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

Submission document: Testing attachment, section 2b2.3

Panel Member #1: Correlation with STAR ratings mortality score 0.422. Most facilities met standard of >60% non-missing data for imputation. Person correlation of score with and without imputed data 0.993 and weighted kapp 0.842. Registry and administrative data: 93% within 5 points, 84% within 2. Levels of correlation are sufficient, although the confidence intervals for RSMR across quartiles of STAR ratings substantially overlap. Imputation justified. Face validity established by expert and stakeholder input.
Panel Member #2: The empirical validity of the NIH stroke scale in terms of its alignment with the GWTG-registry was very good (corr = 0.993 and a weighted kappa of 0.842.). The validation of the RSMR in terms of its correlation with Overall Star Ratings Mortality scores is 0.422, which is low.
Face validity exercise have many missing details. We don't know how the advisory panel was constituted, the composition of the advisory group and their potential conflict of interests. It is also not clear as to what specific ways these members assessed whether the measure can distinguish between good and poor-quality care (e.g., voting on the appropriateness of the measure on a Likert scale?).
Panel Member #4: Data element validity is compelling. Measure score validity methods and results are not compelling. There is no explicit quality construct nor demonstration of a construct-outcome relationship.
Panel Member #5: The measure has been widely used by many researchers in published articles.
Panel Member #6: When comparing the NIH Stoke Scale scores within the GWTG-Stroke Registry and administrative claims data, 93% of the scores from the two data sources were within 5 points of each other and 84% within 2 points. The distributions of NIH Stroke Scale scores from the administrative and GWTG-Stroke Registry data were similar with a Pearson Correlation Coefficient of 0.993 and a weighted kappa of 0.842.
Empirical validity of measure score using the Overall Star Ratings Mortality Measure Group score found a correlation of 0.422.
The face validity testing was reported to have demonstrated working group agreement with face validity of the measure but those **results were not presented**.
Overall, these results support strong data element validity and high measure score validity.
Panel Member #8: True empirical score validity testing was not performed. Validity is based on correlation with other databases, the STAR Ratings or with expert opinion. The alignment with the established Stroke Registry was intentional.
Panel Member #9: Correlation between NIH Stroke Scale scores in ICD-10 claims and GWTG Registry had a Kappa of 0.842 to support NIH scale for risk adjustment. Correlation with Hospital Star Rating demonstrated but question the use of Star Ratings to determine quality of care.

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)

Panel Member #1: but see note in 22.

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)

Panel Member #1: We have previously accepted the argument that CMS auditing of data for payment was an acceptable measure of data element accuracy

Panel Member #9: Only NIH Stroke Scale was analyzed

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

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

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

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

☐ **Insufficient** (NOTE: For instrument-based measures and some composite measures, testing at both the score level and the data element level is required; if not conducted, should rate as INSUFFICIENT.)

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: Measure has face validity and reasonable correlation with Medicare Compare mortality rating.

Panel Member #2: Besides my comments in #21 and #22 above, another reason for my low rating is due to my concern testing sample included only hospitals that reported NIH Stroke scale for least 60% of the ischemic stroke admissions. It is not clear why some hospitals don't report NIH Stroke Scale and how those hospitals will be impacted following the implementation of this score.

Panel Member #4: The rating of low is based on both the method and the result.

Panel Member #5: Developer demonstrated an effort to risk adjust measure to create valid measure score. I have a personal bias against using the RSRR approach (described in S.14) comparing the "predicted" to the "expected" Provider rates because both values are dependent upon the quality (power and specificity) of the regression models. However, the RSRR methodology has been deemed acceptable by SMP by consensus and I will abide by that decision.

Note: The relationship between the Stroke RSRR values and Star Ratings (by quintile) show virtually no discriminatory value among the middle three quintiles based on inter-quartile range and median value. The first and fifth may have a difference of an RSRR value of approximately 1.0. Is this very small difference in RSRR value s meaningful value on which to base payment differences?

Panel Member #6: The validity test results show strong validity of the model based on measure score and data element validity testing.

Panel Member #7: "the Overall Star Rating Mortality measure group scores and the Stroke Mortality with Adjustment for Stroke Severity measure indicate a moderate association, which is to be expected given that these metrics assess different cohorts and the Overall Star Ratings measure group scores are calculated using Latent Variable Modeling, a unique and complex statistical approach in which some of the underlying measures contribute more to the measure group score than other measures. Therefore, the results above show that the trend and results of the Stroke Mortality with Adjustment for Stroke Severity measure are in agreement with other measures of quality."

Panel Member #8: Much of the validity of this measure is due to its correlation with the STAR Ratings and the Stroke Registry Scores. It also has TEP support.

Panel Member #9: No concerns related to NIH Stroke Scale validity. Question use of Star Rating comparison when other outcome measures were available.

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.**

Panel Member #6: See comments related to evaluating the risk adjustment findings related to social risk factors.

Developer Submission

Additional evaluations and submission materials attachments

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_3596_Stroke_Mortality__NQF_Evidence_Attachment_For_NQF_team-637394929740513413.docx

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

1a. Evidence (subcriterion 1a)

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

Measure Title: Hospital 30-day, all-cause, risk-standardized mortality rate (RSMR) following acute ischemic stroke hospitalization with claims-based risk adjustment for stroke severity

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 acute ischemic stroke hospitalization with claims-based risk adjustment for stroke severity

☐ Patient-reported outcome (PRO):

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

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

☐ Process:

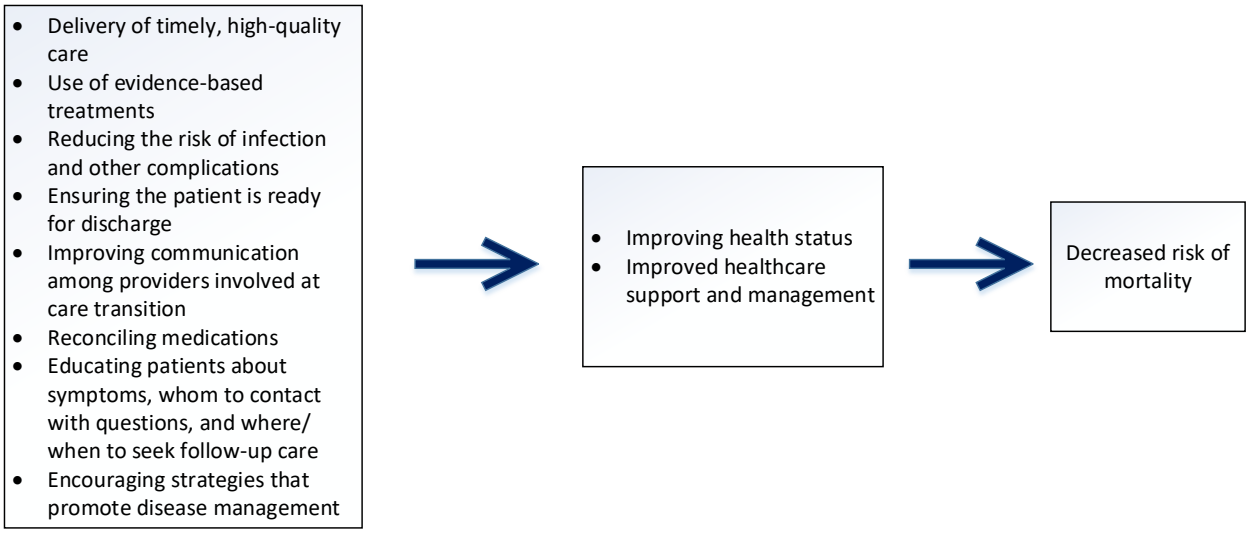
☐ Appropriate use measure:

☐ Structure:

☐ Composite:

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

Figure 1. Stroke Mortality with Adjustment for Stroke Severity Logic Model



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

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

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

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

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

Stroke continues to be a leading cause of mortality and morbidity in the United States, with an estimated 795,000 people having a stroke each year (CDC, 2020). Stroke remains the nation's fifth leading cause of death (CDC, 2020). Among patients 65 and older, both stroke burden and subsequent mortality varies considerably by geographic region (CDC, 2020; Thompson et al., 2017). Most of these strokes are ischemic in nature and increase in prevalence with advancing age (CDC, 2020; Benjamin et al., 2020). Some projections estimated that more than 3 million adults, representing almost 4% of the US adult population, will have had a stroke by 2030 (Ovbiagele et al., 2013). It is estimated that stroke costs $34 billion each year in direct and indirect medical

costs (Mozaffarian et al., 2015; CDC, 2020). As such, stroke mortality is a priority condition for outcomes measure development.

Many current hospital processes have been associated with lower stroke mortality rates within 30 days of hospital admission. In particular, post-stroke mortality rates have been shown to be influenced by critical aspects of care at the hospital such as response to complications, speediness of delivery of care, organization of care, coordinated transitions to the outpatient environment, antihypertensive and anticoagulant therapies, and appropriate imaging (Hacke et al., 2004; Smith et al., 2006; Fang et al., 2008; Reeves et al., 2009; Lingsma et al., 2008; Hong et al., 2008; Fonarow et al., 2014; Bekelis et al., 2016; Xian et al., 2019; Jahan et al., 2019). This research demonstrates the relationship between hospital organizational factors and performance on the acute ischemic stroke mortality measure and supports the ability of hospitals to impact these rates. For example, hospitals participating in quality improvement registries like Get With The Guidelines (GWTG) had lower in-hospital mortality rates among stroke patients than hospitals not participating in similar programs (Fonarow et al., 2014). Another study found that patients being treated at hospitals participating in the GWTG quality improvement registry for stroke were significantly more likely to receive multiple evidence-based care interventions, such as tissue plasminogen activator (tPA) administration and evaluation by a neurologist (Howard et al., 2018). Risk-adjusted measures of patient outcomes, specifically mortality, can highlight variations in the provision of care, and thus support improvements by highlighting institutions that provide exceptional care for stroke patients.

Stakeholders have previously stressed the importance of including stroke severity in mortality measures for risk adjustment, as several studies have demonstrated that initial stroke severity is the strongest predictor of mortality in acute ischemic stroke patients (Smith et al., 2010; Nedeltchev et al., 2010; Fonarow et al., 2012; Lichtman et al., 2019). This update to the current publicly reported measure responds to stakeholder preference to include the National Institutes of Health (NIH) Stroke Scale as an assessment of stroke severity in the risk-adjustment model, thereby accounting for stroke severity at the time of admission to assess the condition of the patient before care has been administered. Moreover, the inclusion of the NIH Stroke Scale has been shown to improve model discrimination for the publicly reported stroke mortality measure (Schwartz et al., 2017).

References:

Bekelis K, Marth NJ, Wong K, Zhou W, Birkmeyer JD, Skinner J. Primary Stroke Center Hospitalization for Elderly Patients With Stroke: Implications for Case Fatality and Travel Times. *JAMA Intern Med.* 2016;176(9):1361-1368.

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.

Centers for Disease Control and Prevention: Stroke. Available at: https://www.cdc.gov/stroke/index.htm. Accessed July 27, 2020.

CDC, NCHS. Underlying Cause of Death 1999-2013 on CDC WONDER Online Database, released 2015. Data are from the Multiple Cause of Death Files, 1999-2013, as compiled from data provided by the 57 vital statistics jurisdictions through the Vital Statistics Cooperative Program. Accessed Feb. 3, 2015.

Fang J, Keenan NL, Ayala C, Dai S, Merritt R, Denny CH. Awareness of stroke warning symptoms—13 states and the District of Columbia, 2005. MMWR. 2008;57(18):481–5.

Fonarow GC, Saver JL, Smith EE, et al. Relationship of national institutes of health stroke scale to 30-day mortality in medicare beneficiaries with acute ischemic stroke. J Am Heart Assoc. Feb 2012;1(1):42-50.

Fonarow GC, Zhao X, Smith EE, et al. Door-to-Needle Times for Tissue Plasminogen Activator Administration and Clinical Outcomes in Acute Ischemic Stroke Before and After a Quality Improvement Initiative. JAMA. 2014;311(16):1632-1640.

Hacke W, Donnan G, Fieschi C, et al. Association of outcome with early stroke treatment: pooled analysis of ATLANTIS, ECASS, and NINDS rt-PA stroke trials. Lancet (London, England). Mar 6 2004;363(9411):768-774.

Hong KS, Kang DW, Koo JS, et al. Impact of neurological and medical complications on 3-month outcomes in acute ischaemic stroke. European journal of neurology: the official journal of the European Federation of Neurological Societies. Dec 2008;15(12):1324-1331.

Howard G, Schwamm LH, Donnelly JP, et al. Participation in Get With The Guidelines-Stroke and Its Association With Quality of Care for Stroke. *JAMA Neurol*. 2018;75(11):1331-1337. doi:10.1001/jamaneurol.2018.2101.

Jahan R, Saver JL, Schwamm LH, et al. Association Between Time to Treatment With Endovascular Reperfusion Therapy and Outcomes in Patients With Acute Ischemic Stroke Treated in Clinical Practice. *JAMA*. 2019;322(3):252-263. doi:10.1001/jama.2019.8286.

Lichtman JH, Leifheit EC, Wang Y, Goldstein LB. Hospital Quality Metrics: "America's Best Hospitals" and Outcomes After Ischemic Stroke. *J Stroke Cerebrovasc Dis.* 2019;28(2):430-434.

Lingsma HF, Dippel DW, Hoeks SE, et al. Variation between hospitals in patient outcome after stroke is only partly explained by differences in quality of care: results from the Netherlands Stroke Survey. Journal of neurology, neurosurgery, and psychiatry. Aug 2008;79(8):888-894.

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

Nedeltchev K, Renz N, Karameshev A, et al. Predictors of early mortality after acute ischaemic stroke. Swiss Medical Weekly. 2010;140(17-18):254-259.

Ovbiagele B, Goldstein LB, Higashida RT, et al. Forecasting the future of stroke in the United States: a policy statement from the American Heart Association and American Stroke Association [published correction appears in Stroke. 2015 Jul;46(7):e179]. *Stroke*. 2013;44(8):2361-2375. doi:10.1161/STR.0b013e31829734f2.

Reeves MJ, Smith E, Fonarow G, Hernandez A, Pan W, Schwamm LH. Off-hour admission and in-hospital stroke case fatality in the get with the guidelines-stroke program. Stroke. Feb 2009;40(2):569-576.

Schwartz J, Wang Y, Qin L, et al. Incorporating Stroke Severity Into Hospital Measures of 30-Day Mortality After Ischemic Stroke Hospitalization. *Stroke.* 2017;48(11):3101-3107.

Smith EE, Shobha N, Dai D, et al. Risk score for in-hospital ischemic stroke mortality derived and validated within the Get With the Guidelines-Stroke Program. Circulation. 2010;122(15):1496-1504.

Smith MA, Liou JI, Frytak JR, Finch MD. 30-day survival and rehospitalization for stroke patients according to physician specialty. Cerebrovascular diseases (Basel, Switzerland). 2006;22(1):21-26.5.

Thompson MP, Zhao X, Bekelis K, et al. Regional Variation in 30-Day Ischemic Stroke Outcomes for Medicare Beneficiaries Treated in Get With The Guidelines-Stroke Hospitals. *Circ Cardiovasc Qual Outcomes*. 2017;10(8):e003604. doi:10.1161/CIRCOUTCOMES.117.003604.

Xian Y, Xu H, O'Brien EC, et al. Clinical Effectiveness of Direct Oral Anticoagulants vs Warfarin in Older Patients With Atrial Fibrillation and Ischemic Stroke: Findings From the Patient-Centered Research Into Outcomes Stroke Patients Prefer and Effectiveness Research (PROSPER) Study [published online ahead of print, 2019 Jul 22].

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.	*
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?	*

Systematic Review	Evidence
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.

Stroke is the fifth most common cause of death, affecting approximately 795,000 people in the United States annually [CDC, 2020], and has a mortality rate of 17% [Go et al., 2014; Kochanek et al., 2014]. Stroke is also a leading cause of disability in the United States, which can lead to increased dependency on the health care system and higher subsequent costs associated with this care [Mozaffarian et al., 2015; CDC, 2020]. Mortality following stroke – an important adverse outcome that can be measured reliably and objectively, and that is influenced by the quality of care provided to patients during their initial hospitalization – is an appropriate measure of quality of care [DesHarnais et al., 1988; Weir et al, 2001]. Specifically, post-stroke mortality rates have been shown to be influenced by critical aspects of care such as response to complications, speediness of delivery of care, organization of care, and appropriate imaging [Hacke et al., 2004; Smith et al., 2006; Fang et

al., 2008; Reeves et al., 2009; Lingsma et al., 2008; Hong et al., 2008; Fonarow et al., 2014; Bekelis et al., 2016; Xian et al., 2019; Jahan et al., 2019]. This work demonstrates the relationship between hospital organizational factors and performance on the stroke mortality measure and supports the ability of hospitals to impact these rates.

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 acute ischemic stroke. 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 outcome 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.

Rationale for Development of an Updated Claims-Only Stroke Mortality Measure

Current outcome measures use administrative claims data from the year prior to the index admission in the risk adjustment models. Stakeholders, including the AHA/ASA and other professional organizations, have highlighted the importance of including stroke severity in mortality measures for risk adjustment. Several studies have demonstrated that initial stroke severity is the strongest predictor of mortality in ischemic stroke patients [Smith et al., 2010; Nedeltchev et al., 2010; Fonarow et al., 2012].

This new claims-based stroke mortality measure addresses these stakeholder preferences and improves model performance by updating the current publicly reported claims-based stroke mortality measure to incorporate stroke severity scores into the risk-adjustment model. Advancements in clinical practice to incorporate new clinical assessments in administrative coding systems have made it possible to integrate these data into measures of hospital performance. The NIH Stroke Scale, which was created in 1989 and is widely used in routine stroke care, is collected in the GWTG-Stroke Registry, which has over 1,700 hospitals throughout the U.S. [Fonarow et al., 2014]. The NIH Stroke Scale is a 15-item neurologic examination stroke scale used to provide a quantitative assessment of stroke related neurologic deficit, by evaluating the effect of acute ischemic stroke on the levels of consciousness, language, neglect, visual-field loss, extraocular movement, motor strength, ataxia, dysarthria, and sensory loss. The NIH Stroke Scale is designed to be a simple, valid, and reliable tool that can be consistently administered at the bedside by physicians, nurses, or therapists. The use of the NIH Stroke Scale to assess stroke severity upon acute ischemic stroke patient presentation is recommended in the AHA/ASA Class I guidelines [Powers et al, 2019]. Furthermore, as of October 2016, the NIH Stroke Scale score is now coded as a secondary ICD-10-CM code within administrative claims, allowing it to be used in this measure. Inclusion of stroke severity data will not only address stakeholder preferences, but also improve the discrimination of the risk model.

References:

Adams HP, Jr., Davis PH, Leira EC, et al. Baseline NIH Stroke Scale score strongly predicts outcome after stroke: A report of the Trial of Org 10172 in Acute Stroke Treatment (TOAST).

Bekelis K, Marth NJ, Wong K, Zhou W, Birkmeyer JD, Skinner J. Primary Stroke Center Hospitalization for Elderly Patients with Stroke: Implications for Case Fatality and Travel Times. JAMA Intern Med. 2016;176(9):1361-1368.

Casper ML NI, Croft JB, Nilasena DS. Atlas of Stroke Hospitalizations Among Medicare Beneficiaries. 2008.

Centers for Disease Control and Prevention: Stroke. Available at: https://www.cdc.gov/stroke/index.htm. Accessed July 27, 2020.

DesHarnais SI, Chesney JD, Wroblewski RT, Fleming ST, McMahon LF, Jr. The Risk-Adjusted Mortality Index. A new measure of hospital performance. Med Care. Dec 1988;26(12):1129-1148.

Fang J, Keenan NL, Ayala C, Dai S, Merritt R, Denny CH. Awareness of stroke warning symptoms—13 states and the District of Columbia, 2005. MMWR. 2008;57(18):481–5.

Fonarow GC, Alberts MJ, Broderick JP, et al. Stroke outcomes measures must be appropriately risk adjusted to ensure quality care of patients: a presidential advisory from the American Heart Association/American Stroke Association. Stroke. May 2014;45(5):1589-1601.

Fonarow GC, Saver JL, Smith EE, et al. Relationship of national institutes of health stroke scale to 30-day mortality in Medicare beneficiaries with acute ischemic stroke. J Am Heart Assoc. Feb 2012;1(1):42-50.

Go AS, Mozaffarian D, Roger VL, et al. Heart disease and stroke statistics--2014 update: a report from the American Heart Association. Circulation. Jan 21, 2014;129(3): e28-e292.

Hacke W, Donnan G, Fieschi C, et al. Association of outcome with early stroke treatment: pooled analysis of ATLANTIS, ECASS, and NINDS rt-PA stroke trials. Lancet (London, England). Mar 6 2004;363(9411):768-774.

Hong KS, Kang DW, Koo JS, et al. Impact of neurological and medical complications on 3-month outcomes in acute ischaemic stroke. European journal of neurology: the official journal of the European Federation of Neurological Societies. Dec 2008;15(12):1324-1331.

Jahan R, Saver JL, Schwamm LH, et al. Association Between Time to Treatment with Endovascular Reperfusion Therapy and Outcomes in Patients with Acute Ischemic Stroke Treated in Clinical Practice. JAMA. 2019;322(3):252-263. doi:10.1001/jama.2019.8286.

Kochanek KD MS, Xu JQ, Arias E. Mortality in the United States, 2013. NCHS data brief, no. 178. 2014.

Lingsma HF, Dippel DW, Hoeks SE, et al. Variation between hospitals in patient outcome after stroke is only partly explained by differences in quality of care: results from the Netherlands Stroke Survey. Journal of neurology, neurosurgery, and psychiatry. Aug 2008;79(8):888-894.

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

Nedeltchev K, Renz N, Karameshev A, et al. Predictors of early mortality after acute ischaemic stroke. Swiss Medical Weekly. 2010;140(17-18):254-259.

Powers WJ, Rabinstein AA, Ackerson T, et al. Guidelines for the Early Management of Patients with Acute Ischemic Stroke: 2019 Update to the 2018 Guidelines for the Early Management of Acute Ischemic Stroke: A Guideline for Healthcare Professionals from the American Heart Association/American Stroke Association. Stroke. 2019;50(12): e344-e418.

Reeves MJ, Smith E, Fonarow G, Hernandez A, Pan W, Schwamm LH. Off-hour admission and in-hospital stroke case fatality in the get with the guidelines-stroke program. Stroke. Feb 2009;40(2):569-576.

Smith EE, Shobha N, Dai D, et al. Risk score for in-hospital ischemic stroke mortality derived and validated within the Get with the Guidelines-Stroke Program. Circulation. 2010;122(15):1496-1504.

Smith MA, Liou JI, Frytak JR, Finch MD. 30-day survival and rehospitalization for stroke patients according to physician specialty. Cerebrovascular diseases (Basel, Switzerland). 2006;22(1):21-26.

Weir NU, Sandercock PA, Lewis SC, Signorini DF, Warlow CP. Variations between countries in outcome after stroke in the International Stroke Trial (IST). Stroke. Jun 2001;32(6):1370-1377.

Xian Y, Xu H, O´Brien EC, et al. Clinical Effectiveness of Direct Oral Anticoagulants vs Warfarin in Older Patients with Atrial Fibrillation and Ischemic Stroke: Findings from the Patient-Centered Research into Outcomes Stroke Patients Prefer and Effectiveness Research (PROSPER) Study [published online ahead of print, 2019 Jul 22].

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.*

For testing purposes only, we used Medicare Fee-For-Service administrative claims data from October 1, 2016 to June 30, 2019, for hospitals where the National Institute of Health (NIH) Stroke Scale was coded on 60% of claims. Our cohort included 89,795 patients at 329 hospitals. The mean risk-standardized mortality rate (RSMR) among hospitals was 14.63% and the median hospital RSMR was 14.68%, with a range of 10.05% to 17.83% and an interquartile range of 13.82% to 15.52%.

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

Mortality following acute ischemic stroke is an important adverse outcome that can be measured reliably and objectively. The 30-day stroke mortality rate varies by age from 9% in patients 65-74 years of age to 23% in those =85 years of age [Casper et al., 2008]. Post-stroke mortality rates have also been shown to be influenced by critical aspects of care.

Risk-adjusted measures of patient outcomes, including mortality, can highlight variation in the care patients receive across hospitals, and thus support improvements and learning from high quality institutions.

The results of CMS's current publicly reported claims-based stroke mortality measure are based on RSMRs calculated for admissions among Medicare FFS patients, age 65 years and older. Risk-adjusted 30-day mortality rates were shown to decline over the three-year period of July 2013 through June 2016 from 14.9% (between July 2013 and June 2014) to 14.0% (between July 2015 and June 2016). The median hospital RSMR in the combined three-year dataset was 14.5% (IQR: 14.0% - 15.3%) [CMS Hospital Chartbook].

Between July 2016-June 2017 and July 2018-June 2019, the observed rate decreased from 13.9% to 13.3%. The median hospital RSMR in the combined three-year dataset was 13.6% (IQR: 13.0% - 14.2%).

This dataset includes 520,432 admissions from 4,254 hospitals. This decline suggests that there is opportunity for further improvement in the 30-day mortality outcome over time.

References:

Casper ML NI, Croft JB, Nilasena DS. Atlas of Stroke Hospitalizations Among Medicare Beneficiaries. 2008.

CMS Hospital Chartbook: Trends in mortality rates following admission for acute myocardial infarction, chronic obstructive pulmonary disease, heart failure, pneumonia, and acute ischemic stroke. https://www.cmshospitalchartbook.com/stroke-mortality. Accessed October 2, 2020.

Simoes J, Grady JN, Debuhr J, et al. 2017 Condition-Specific Measures Updates and Specifications Report Hospital-Level 30-Day Risk-Standardized Mortality Measures: Acute Myocardial Infarction – Version 11.0 Chronic Obstructive Pulmonary Disease – Version 6.0 Heart Failure – Version 11.0 Pneumonia – Version 11.0 Stroke – Version 6.0. https://www.qualitynet.org/inpatient/measures/mortality/resources#tab3. Available as of March 16, 2017.

Debuhr J, McDowell K, Grady JN, et al. 2020 Condition-Specific Measures Updates and Specifications Report Hospital-Level 30-Day Risk-Standardized Mortality Measures: Acute Myocardial Infarction – Version 14.0 Chronic Obstructive Pulmonary Disease – Version 9.0 Heart Failure – Version 14.0 Pneumonia – Version 14.0 Stroke – Version 9.0. https://www.qualitynet.org/inpatient/measures/mortality/methodology. Available as of May 20, 2020.

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 claims-only stroke RSMRs by proportion of Dual Eligible patients:

Data source: Medicare FFS claims and Master Beneficiary Summary File (MBSF) data

Dates of data: October 1, 2016 – June 30, 2019

Characteristic//Hospitals in the bottom quartile of proportion Dual Eligible patients//Hospitals in the top quartile of proportion Dual Eligible patients –

Number of Measured Entities (Hospitals)// 83 // 82

Percentage of dual eligible patients // <=7.73 // >=16.90

Minimum // 11.72 // 10.05

10th percentile// 13.50 // 12.84

25th percentile// 14.05 // 13.90

Median (50th percentile)// 14.53 // 14.33

75th percentile// 15.04// 14.82

90th percentile// 15.79 // 15.25

Maximum// 17.21 // 17.44

Distribution of claims-only stroke RSMRs by AHRQ SES Index:

Data source: Medicare FFS claims and the American Community Survey (2008-2012) data

Dates of Data: October 1, 2016 – June 30, 2019 (claims); 2013-2017 (ACS)

Characteristic//Hospitals in bottom quartile of AHRQ SES Index //Hospitals in top quartile of AHRQ SES Index –

Number of Measured Entities (Hospitals)// 81 // 82

Percentage of dual eligible patients // <=5.22 // >=23.12

Minimum // 12.34 // 11.76

10th percentile// 13.79 // 13.23

25th percentile// 14.14 // 13.86

Median (50th percentile)// 14.52 // 14.51

75th percentile// 14.82 // 15.32

90th percentile// 15.45 // 16.28

Maximum// 16.16 // 17.83

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):*

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

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

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.)*

This measure adds risk adjustment for stroke severity to an existing stroke mortality measure reported in the Inpatient Quality Reporting (IQR) program. The outcome, cohort, and measure calculation are the same for both measures and can be reviewed on Qua

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**:** Stroke_Mortality_w.NIHSS_datadictionary_Final-637320587861997683.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.

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

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 of the index admission for Medicare FFS patients aged 65 years and older with a principal discharge diagnosis of acute ischemic stroke.

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 deaths for any cause within 30 days of the index acute ischemic stroke admission. As currently specified, we identify deaths for FFS Medicare patients 65 years and older in the Medicare Enrollment Database (EDB).

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

The cohort includes inpatient admissions to all non-federal, short-term, acute care or critical access hospitals for Medicare FFS patients aged 65 years and older with a principal discharge diagnosis of acute ischemic stroke.

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 additional inclusion criteria:

1. Principal discharge diagnosis of acute ischemic stroke
2. Enrolled in Part A and Part B Medicare for the 12 months prior to the date of index admission, and Medicare FFS during the index admission
3. Aged 65 or over
4. Not transferred from another acute care facility

A list of ICD-10 codes that define the patient cohort are included in the Data Dictionary.

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

The mortality measure excludes index admissions for patients:

1. With inconsistent or unknown vital status or other unreliable data;
2. Enrolled in the Medicare hospice program at any time in the 12 months prior to the index admission, including the first day of the index admission; and
3. Discharged against medical advice (AMA).

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

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

1. 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.

2. 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.

3. 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.

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 stroke 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, 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 on the basis of 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 to the sum of the estimated regression coefficients multiplied by the patient characteristics. The results are transformed and summed over all patients attributed to a hospital to get a predicted value. The "expected" number of deaths (the denominator) is obtained in the same manner, but a common intercept using all hospitals in 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 mortality 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:

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

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

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

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

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

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

N/A

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

If other, please describe in S.18.

Claims, Enrollment Data, Other, Registry Data

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

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

For measure specification and testing the data sources were:

Medicare Part A inpatient and Part B outpatient claims: This data source contains claims data for fee-for service 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 from the EDB that contains enrollment information for all Medicare beneficiaries including dual-eligible status. Years 2016-2019 were used.

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.

Overall Hospital Quality Star Ratings Mortality Measure Group: This data contains a summary of mortality measures, using October 2019 Hospital Compare data. This data was used to test measure score validity.

American Heart Association/American Stroke Association (AHA/ASA)'s Get With The Guidelines (GWTG)-Stroke Registry: This data contains NIH Stroke Scale scores derived from patient medical records from 2016-2019. This data was used to test data element validity.

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. Data sources for the all-payer update

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

Validity – See attached Measure Testing Submission Form

Stroke_Mortality_w.NIHSS_NQF_Testing_form_Final_updated_10.28.20.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.

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.

2.3 For maintenance of endorsement

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

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

Measure Number (*if previously endorsed*)**:**
Measure Title: Hospital 30-day, all-cause, risk-standardized mortality rate (RSMR) following acute ischemic stroke hospitalization with claims-based risk adjustment for stroke severity
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

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

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

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

AND

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

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

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

OR

- rationale/data support no risk adjustment/stratification.

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

OR

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

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

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

Notes

10. Reliability testing applies to both the data elements and computed measure score. Examples of reliability testing for data elements include, but are not limited to: inter-rater/abstractor or intra-rater/abstractor

studies; internal consistency for multi-item scales; test-retest for survey items. Reliability testing of the measure score addresses precision of measurement (e.g., signal-to-noise).

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

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

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

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

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

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

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

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

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

1.2. If an existing dataset was used, identify the specific dataset (*the dataset used for testing must be consistent with the measure specifications for target population and healthcare entities being measured; e.g.,*

Medicare Part A claims, Medicaid claims, other commercial insurance, nursing home MDS, home health OASIS, clinical registry).

The dataset used for testing included Medicare Parts A and B claims as well as the Medicare Enrollment Database (EDB). Data from the American Heart Association/American Stroke Association (AHA/ASA)'s Get With The Guidelines (GWTG)-Stroke Registry, as well as from the Overall Hospital Quality Star Ratings Mortality measure group were used for measure validation. 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 score derived from census data). The dataset used varies by testing type; see Section 1.7 for details.

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

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

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

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

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

For this measure, hospitals are the measured entities. The testing dataset included 329 non-federal, acute inpatient US hospitals that submitted claims with National Institute of Health (NIH) Stroke Scale scores for at least 60% of their stroke admissions. The number of measured entities varies slightly by the type of testing performed; 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.

For measure testing, we used Medicare administrative claims data from October 1, 2016, when International Classification of Disease, Tenth Revision (ICD-10) data became available, to June 30, 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. We randomly split the data into two equal samples: Development Dataset and Internal Validation Dataset.

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	Dates of Data: October 1, 2016 – June 30, 2019 Number of admissions = 89,795 Number of measured hospitals: 329 This cohort was randomly split for initial model testing. First half of split sample (Development) ***Number of Admissions: 44,898*** ***Number of Measured Hospitals: 326*** Second half of split sample (Validation) ***Number of Admissions: 44,897*** ***Number of Measured Hospitals: 322***
Testing Dataset **(Medicare Fee-For-Service Administrative Claims Data)**	Section 2a2 Reliability Testing Section 2b1 Validity Testing Section 2b2 Testing of Measure Exclusion Section 2b3 Risk Adjustment/Stratification 2b3.6. Statistical Risk Model Discrimination Statistics Section 2b4 Meaningful Differences	Dates of Data: October 1, 2016 – June 30, 2019 Number of admissions = 89,795 Patient Descriptive Characteristics: mean age = 79.5 years; % male = 44.8 Number of measured hospitals: 329
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 social risk factors. 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: October 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.
American Heart Association/American Stroke Association (AHA/ASA)'s Get With The Guidelines (GWTG)-Stroke Registry	Section 2b1: Validity Testing	Dates of Data: October 1, 2016 – December 31, 2019 We used NIH Stroke Scale scores from AHA/ASA's GWTG Stroke Registry data to validate the NIH Stroke Scale scores coded within administrative claims.
Overall Hospital Quality Star Ratings Mortality Measure Group	Section 2b1: Validity Testing	Dates of Data: January 2020 release of the Overall Star Ratings, using October 2019 *Hospital Compare* data The Mortality measure group consists of 7 measures. The dates of data for each of the measures is as follows: • MORT-30-AMI: July 1, 2015 – June 30, 2018 • MORT-30-CABG: July 1, 2015 – June 30, 2018 • MORT-30-COPD: July 1, 2015 – June 30, 2018 • MORT-30-HF: July 1, 2015 – June 30, 2018 • MORT-30-PN: July 1, 2015 – June 30, 2018 • MORT-30-STK: July 1, 2015 – June 30, 2018 • PSI-4-SURG-COMP: July 1, 2016 – June 30, 2018 We used data from the Overall Star Ratings Mortality measure group scores derived from January 2020 Overall Star Ratings, which uses October 2019 *Hospital Compare* data, to study the association between Mortality group scores and the 30-day measure RSMR.

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 commonly 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 for 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 the Department of Health and Human Services Assistant Secretary for Policy and Evaluation (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 takes into account 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 socioeconomic index based on census data linked to census block, the smallest geographic unit possible.

- AHRQ-validated SES index score (summarizing the information from the following 7 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 SES 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, 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 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 (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.

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.

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.

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 estimated the signal to noise reliability (facility-level reliability), which is the reliability with which individual units (hospitals) are measured. The reliability of any one facility's measure score will vary depending on the number of patients admitted for ischemic stroke. 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 the 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 measure score would be calculated using all hospitals but would only be publicly reported for hospitals with at least 25 cases. Therefore, the facility-level reliability was tested with all hospitals and presented for both all hospitals and hospital least 25 admissions that would receive a publicly reported score.

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.

Reference:

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

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

There were 89,795 admissions in the testing sample. We calculated the signal to noise reliability score for all hospitals in the testing sample (N=329) and hospitals with at least 25 cases (N=292) to isolate the hospitals with enough ischemic stroke admissions to receive a publicly reported RSMR (see Figure 1 and Figure 2 below). For all hospitals in the testing sample, the median reliability score was 0.72, ranging from 0.01 to 0.95. The 25th and 75th percentiles were 0.51 and 0.83, respectively. The median reliability score demonstrates sufficient reliability. For hospitals with at least 25 cases, the median reliability score was 0.75, ranging from 0.24 to 0.95. The 25th and 75th percentiles were 0.59 and 0.83, respectively. The median reliability score demonstrates sufficient reliability.

Figure 1. Signal to noise reliability distribution for Stroke Mortality with Adjustment for Stroke Severity for all hospitals (N=329)

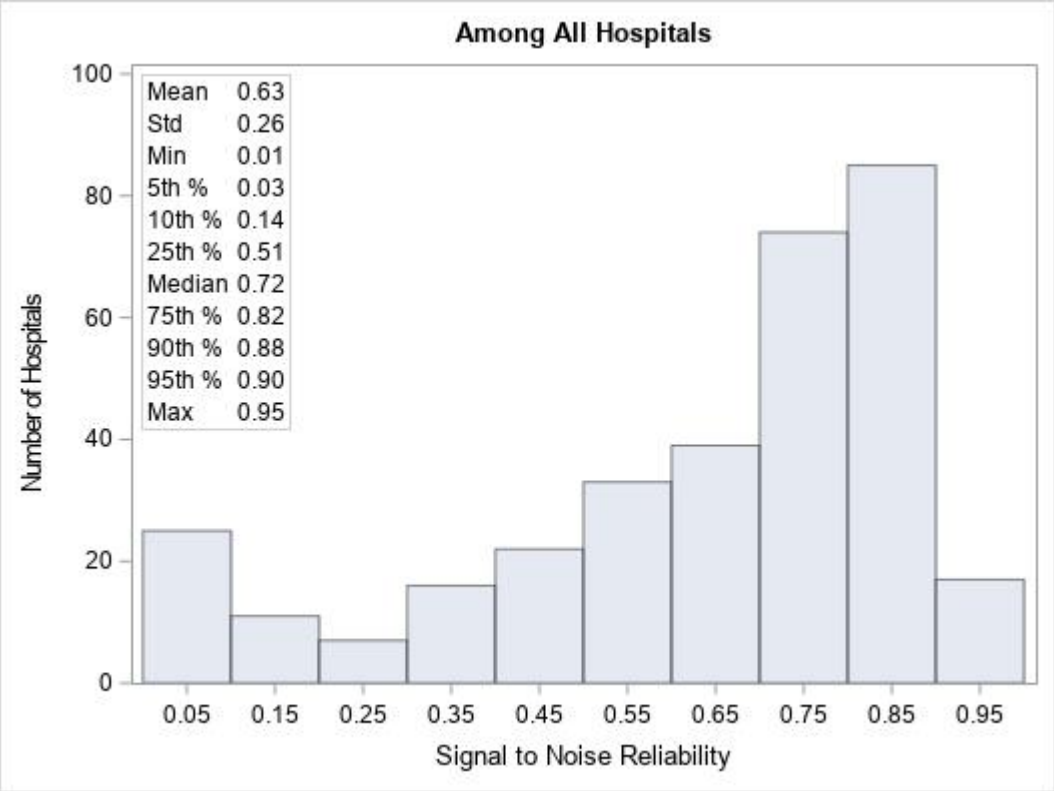
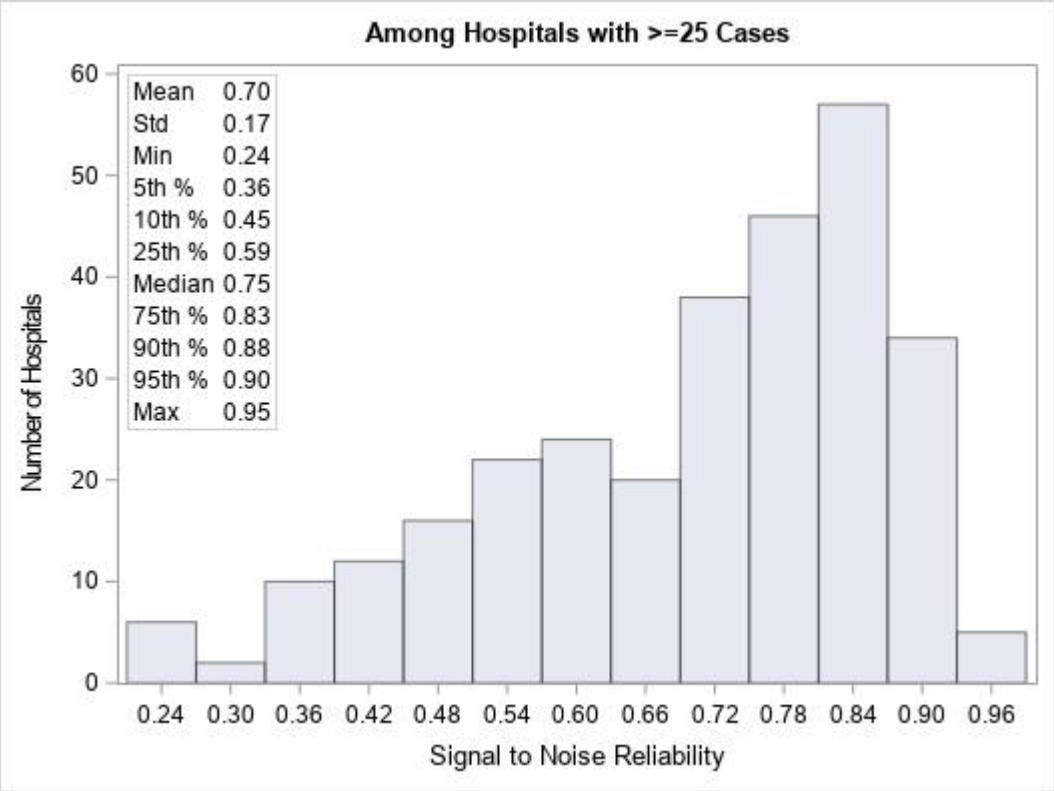


Figure 2. Signal to noise reliability distribution for Stroke Mortality with Adjustment for Stroke Severity for all hospitals (N=292)



Reference:

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

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

Using the approach used by Adams et. al., we obtained the median signal-to-noise reliability score of 0.75, which demonstrates sufficiently high reliability (Adams et al, 2010).

Reference:

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

2b1. VALIDITY TESTING

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

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

☒ **Performance measure score**

☒ **Empirical validity testing**

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

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

Data Element Validity using GWTG-Stroke Registry

We focused data element validity on the NIH Stroke Scale, as the focus of the re-specification of this measure. To assess data element validity of the NIH Stroke Scale scores coded in the claims data, we linked the Medicare claims with the GWTG-Stroke Registry data derived from patients' medical records and compared the scores. The GWTG-Stroke Registry draws data from medical records and has been shown to be reliable through the studies comparing registry data to chart abstraction (Xian et al., 2012). Because only patients aged 65 years and older were included, and some data were excluded based on linkage and other factors, a total of 29,937 stroke hospitalizations were used in the analysis.

Of the linked stroke hospitalizations for which claims data had a non-missing NIH Stroke Scale, we compared the scores recorded in the claims data to the scores in the GWTG-Stroke Registry data. We also examined the distribution of the stroke severity scales within the claims data and the registry data.

Empirical Validity of Measure Score (RSMR) using Overall Star Ratings Mortality Measure Group Score

To further test empirical validity for the Stroke Mortality with Adjustment for Stroke Severity measure, we identified and assessed the measure score's correlation with the Overall Hospital Quality Star Ratings Mortality Measure Group Score, which targets the same domain of quality (mortality) for similar populations. The goal was to identify if better performance on the Stroke Mortality with Adjustment for Stroke Severity measure was related to better performance on the Overall Star Ratings Mortality measure group.

CMS' Overall Hospital Quality Star Ratings assesses hospitals' overall performance (expressed on *Hospital Compare* graphically, as stars) based on a weighted average of group scores from seven measure groups, including a mortality measure group. The mortality group is comprised of the mortality measures that are publicly reported on *Hospital Compare*, including the Hospital 30-Day Mortality Following Acute Ischemic Stroke Hospitalization measure currently reported within the Inpatient Quality Reporting program (IQR). 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 321 Medicare FFS hospitals from January 2020. The full methodology for the Overall Hospital Star Rating can be found at: https://www.qualitynet.org/dcs/ContentServer?c=Page&pagename=QnetPublic%2FPage%2FQnetTier3&cid=1228775957165

We examined the relationship of performance on the stroke mortality measure scores (RSMR) to performance quintiles on the Overall Star Ratings Mortality measure group score.

Validity Indicated by Established Measure Development Guidelines:

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

Validity as Assessed by External Groups:

The Stroke Mortality with Adjustment for Stroke Severity measure was developed as an improvement to the existing Hospital 30-Day Mortality Following Acute Ischemic Stroke Hospitalization measure, currently reported within the Inpatient Quality Reporting (IQR) program. Stakeholders expressed concern that the currently implemented measure did not adjust for stroke severity, so the Yale New Haven Health Services Corporation/Center for Outcomes Research and Evaluation (YNHHSC/CORE) developed a stroke mortality measure which incorporated the NIH Stroke Scale to address these stakeholder concerns.

During 2015-2016 measure development, we obtained expert and stakeholder input via regular discussions with an advisory working group and a 30-day public comment period in order to increase transparency and to gain broader input into the measure.

The working group was assembled, and regular meetings were held throughout the development phase. The working group was tailored for development of this measure and consisted of clinicians (neurologists and cardiologists) and other professionals with expertise in biostatistics, measure methodology, and quality improvement. The working group meetings addressed key issues related to measure development, including the deliberation and finalization of key decisions (e.g., defining the measure cohort and outcome) to ensure the measure is meaningful, useful, and well-designed. The working group provided a forum for focused expert review and discussion of technical issues during measure development.

Following completion of the preliminary model, we solicited public comment on the measure. The public comments were then posted publicly for 30 days. The resulting input was taken into consideration during the final stages of measure development and contributed to minor modifications to the measure.

References:

Reeves MJ, Fonarow GC, Xu H, et al. Is Risk-Standardized In-Hospital Stroke Mortality an Adequate Proxy for Risk-Standardized 30-Day Stroke Mortality Data? Findings From Get With The Guidelines-Stroke. Circ Cardiovasc Qual Outcomes. 2017;10(10).

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;113(3):456-462.

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 July 30, 2020.

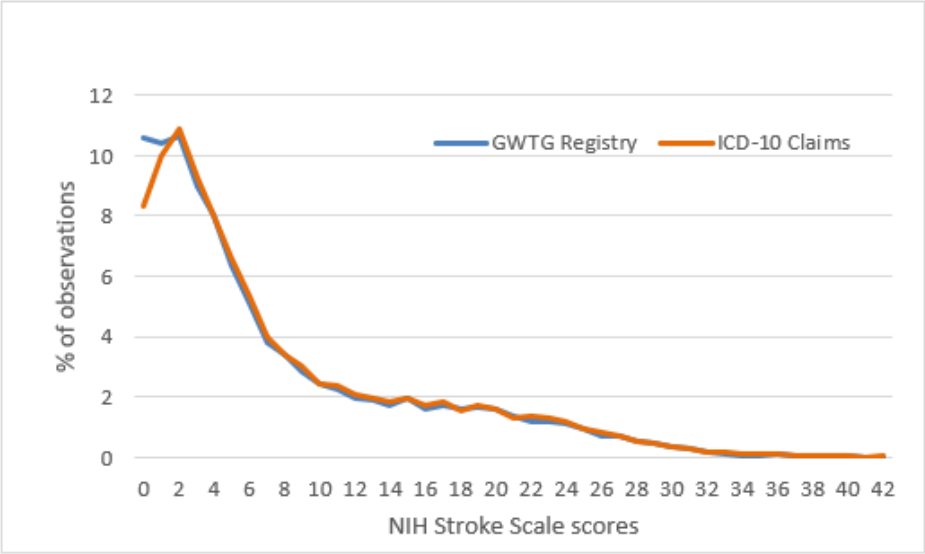
Xian Y, Fonarow GC, Reeves MJ, et al. Data quality in the American Heart Association Get With The Guidelines-Stroke (GWTG-Stroke): Results from a National Data Validation Audit. *American Heart Journal.* 2012;163(3):392-398.e391. http://www.ahjonline.com/article/S0002-8703%2811%2900894-5/abstract

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

Data Element Validity using GWTG-Stroke Registry

When comparing the NIH Stoke Scale scores within the GWTG-Stroke Registry and administrative claims data, 93% of the scores from the two data sources are within 5 points of each other and 84% are within 2 points. The distributions of NIH Stroke Scale scores from the administrative and GWTG-Stroke Registry data are similar with a Pearson Correlation Coefficients: 0.993 and a weighted kappa of 0.842. (See Figure 3 below).

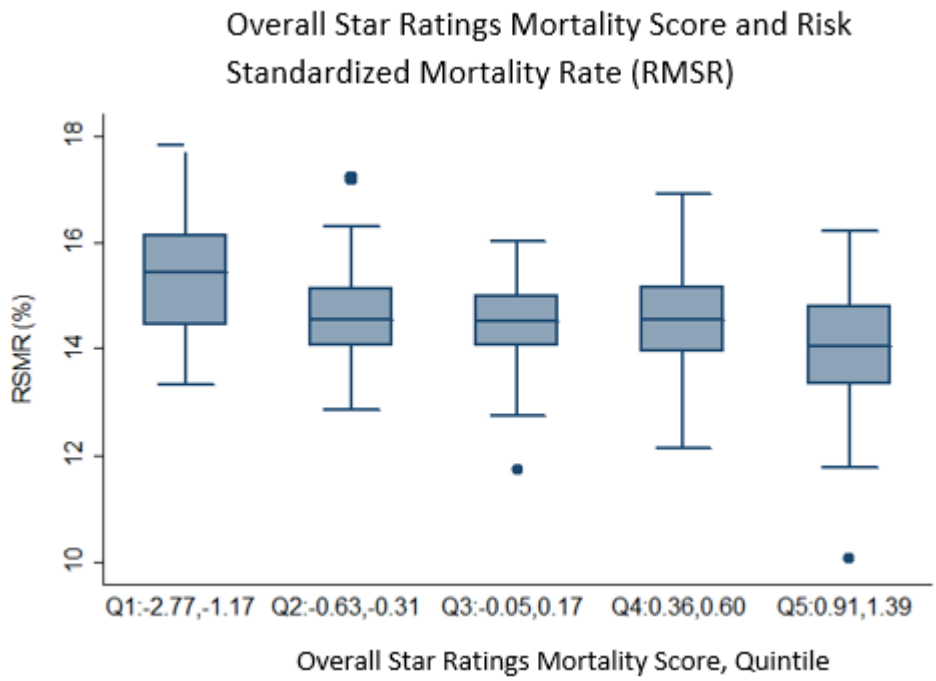
Figure 3. Correlation between NIH Stroke Scale scores in ICD-10 claims and GWTG Registry



Empirical Validity of Measure Score (RSMR) using Overall Star Ratings Mortality Measure Group Score

Figure 4 shows the box-whisker plots of the Stroke Mortality with Adjustment for Stroke Severity RSMRs within each quintiles of Overall Star Ratings Mortality measure group scores. In general, hospitals with higher RSMRs have lower Overall Star Ratings Mortality measure group scores and vice versa. The correlation between stroke RSMRs and the Overall Star Ratings Mortality scores is 0.422.

Figure 4. Box-whisker plots comparing stroke RSMRs and Overall Star Ratings Mortality measure group 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?*)

Data Element Validity using GWTG-Stroke Registry

There was substantial concordance between the NIH Stroke Scale scores within the GWTG-Stroke Registry and administrative claims data, as demonstrated by the close proximity of scores, as well as the Pearson

correlation coefficient of 0.993 and weighted kappa of 0.842. When compared to the GWTG-Stroke Registry NIH Stroke Scale scores, which have been validated through comparison to chart abstraction, the NIH Stroke Scale scores coded on administrative claims can be considered reliable data elements for the adjustment of stroke severity of patients upon admission within the measure.

Empirical Validity Testing

This validation approach compares the Stroke Mortality with Adjustment for Stroke Severity measure results against the Overall Star Rating mortality group scores. Figure 4 Box-whisker plots results demonstrate an observed trend of lower risk-standardized mortality with higher Overall Star Ratings Mortality measure group scores, which supports measure score validity. The results above show that the Stroke Mortality with Adjustment for Stroke Severity measure agrees with external measures of quality. There is a trend in the expected direction providing external support for measure score validity. Also, the correlation coefficients between the Overall Star Rating Mortality measure group scores and the Stroke Mortality with Adjustment for Stroke Severity measure indicate a moderate association, which is to be expected given that these metrics assess different cohorts and the Overall Star Ratings measure group scores are calculated using Latent Variable Modeling, a unique and complex statistical approach in which some of the underlying measures contribute more to the measure group score than other measures. Therefore, the results above show that the trend and results of the Stroke Mortality with Adjustment for Stroke Severity measure are in agreement with other measures of quality.

Validity as Assessed by External Groups

The face validity testing results demonstrated working group agreement with overall face validity of the measure as specified.

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 were made based on clinically relevant decisions and to ensure accurate calculation of the measure. To ascertain impact of exclusions on the cohort, we examined overall frequencies and proportions of the total cohort excluded for each exclusion criterion. These exclusions are consistent with similar NQF-endorsed outcome measures. Rationales for the exclusions are detailed in data field S.10 (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*)

Below is the distribution of exclusions among hospitals:

Table 2. Exclusions Among Hospitals

Exclusion	N	%	Distribution across hospitals (N=329): Min, 25th, 50th, 75th percentile, max
1. Inconsistent or unknown vital status or other unreliable data	2	0.0%	(0.00, 0.00, 0.00, 0.00, 0.00)
2. Enrolled in the Medicare hospice program at any time in the 12 months prior to the index admission, including the first day of the index admission	766	0.70%	(0.00, 0.00, 0.00, 0.01, 0.25)
3. Discharged against medical advice (AMA)	312	0.29%	(0.00, 0.00, 0.00, 0.00, 0.04)

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*)

Exclusions 1 and 2 are necessary for valid calculation of the measure.

Exclusion 1 (patients with inconsistent or unknown vital status or other unreliable demographic [age and gender] data) accounts for 0.0% 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 2 (patients enrolled in the Medicare hospice program any time in the 12 months prior to the index admission, including the first day of the index admission) accounts for 0.70% of all index admissions excluded from the initial index cohort. These patients are likely continuing to seek comfort measures only; mortality is not necessarily an adverse outcome or signal of poor quality care.

Exclusion 3 (patients who are discharged AMA) accounts for 0.29% 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, in the future, 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 would include only the June admission. The July admissions would be 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 20 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 stroke mortality in the 30 days following an index admission. First, we identified the comorbidity or clinical status risk factors that were most important in predicting the outcome, then consider the potential addition of social risk factors.

Our approach to risk adjustment was tailored to and appropriate for a publicly reported outcome measure, as articulated in the American Heart Association (AHA) Scientific Statement, "Standards for Statistical Models Used for Public Reporting of Health Outcomes" (Krumholz et al., 2006).

The measure employs a hierarchical logistic regression model (a form of hierarchical generalized linear model [HGLM]) to create a hospital-level 30-day RSMR. This approach to modeling appropriately accounts for the structure of the data (patients clustered within hospitals), the underlying risk due to patients' comorbidities, and sample size at a given hospital when estimating hospital mortality rates. In brief, the approach simultaneously models two levels (patient and hospital) to account for the variance in patient outcomes within and between hospitals (Normand and Shahian et al., 2007). At the patient level, each model adjusts the log-odds of mortality within 30-days of admission for age, selected clinical covariates and a hospital-specific intercept. The second level models the hospital-specific intercepts as arising from a normal distribution. The hospital intercept, or hospital-specific effect, represents the hospital contribution to the risk of mortality, after accounting for patient risk and sample size, and can be inferred as a measure of quality. The hospital-specific intercepts are given a distribution in order to account for the clustering (non-independence) of patients within the same hospital. If there were no differences among hospitals, then after adjusting for patient risk, the hospital intercepts should be identical across all hospitals.

Clinical Factors

We sought to develop a model that included key variables that were clinically relevant and based on strong relationships with the mortality outcome. We also sought to develop a model that was parsimonious, using a grouper that is in the public domain for the 70,000+ ICD-10-CM codes. The candidate variables for the model were derived from: the index admission, with comorbidities identified from the index admission secondary diagnoses (excluding potential complications); 12-month pre-index inpatient data (for any condition); outpatient hospital data; and Part B physician data. We developed candidate variables for the model from the claims codes.

For risk model development, we started with Condition Categories (CCs) which are part of CMS's Hierarchical Condition Category (HCC). 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. 2000; 2011). These CCs are used in models to predict medical care utilization, mortality, or other related measures.

To select candidate variables for the original Hospital 30-Day Mortality Following Acute Ischemic Stroke Hospitalization Measure, 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.

To re-select candidate variables for this measure and add stroke severity risk adjustment, we began with the list of 42 administrative claims-based risk-adjustment variables included in the current publicly reported Hospital 30-Day Mortality Following Acute Ischemic Stroke Hospitalization Measure. To this, we added the NIH Stroke Scale score from ICD-10 administrative claims for stroke severity. Our set of candidate variables from the claims included 2 demographic variables (age and gender), 39 CC-based variables, an indicator variable for when a patient transferred into the hospital from the emergency department, and the NIH Stroke Scale ICD-10 codes. The final risk-adjustment variables were selected by a team of clinicians and analysts primarily based on their clinical relevance but with knowledge of their strength of association with the mortality outcome.

To develop the model, we began with the 43 candidate predictors for 30-day mortality and followed the same approach of selecting risk variables by using the logistic regression model with the stepwise selection method based on 1,000 bootstrapping samples. Once again, risk variables were retained if they were significantly associated with mortality ($p<0.01$) in 90 percent of the 1,000 repeated samples.

This resulted in a final risk-adjustment model that included 20 variables, including the NIH Stroke Scale score (see Section 2b4.4a table of candidate variables).

Social Risk Factors

We consider adjustment for social risk factors (SRF) 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 ASPE.

In cases where measure results suggest a substantial contribution of a provider effect to the impact of SRFs on a measure outcome, CMS prefers to not include patient-level social risk factors in outcome measure risk-adjustment models.

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, LaPar et al., 2010; Buntin et al., 2017; 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., 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).

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 wealt\h, 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

Statistical Methods

We assessed the relationship between the social risk factor variables with the outcome and examined the incremental effect in a multivariable model. For this measure, we also examined the extent to which the addition of any one of these variables improved model performance or changed hospital results.

References:

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.

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.

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.

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: 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;113(3):456-462.

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

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.

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.

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.

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

☒ **Published literature**

☒ **Internal data analysis**

☐ **Other (please describe)**

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

Below are tables showing the candidate variables and the final variables that were included more than 90% of the time for all copies of the imputed data and therefore retained in the final model with associated odds ratios (OR) and 95 percent confidence intervals (CI). In addition, below is a figure that graphically represents the contribution of each of the risk variables within the final model. Figure 5 demonstrates the importance of the NIH Stroke Scale in predicting mortality for ischemic stroke patients. For every 5-point increase in the NIH Stroke Scale there is a two-fold increase in the odds of mortality.

Table 3. Candidate variables from original Hospital 30-Day Mortality Following Acute Ischemic Stroke Hospitalization measure with addition of NIH Stroke Scale

Variable	Code(s)
Age minus 65 (years above 65, continuous)	*
Male	*
Transfer from another ED	*
Congestive heart failure	CC 85
Valvular or rheumatic heart disease	CC 91
Congenital cardiac/circulatory defects	CC 92, 93
Hypertensive heart disease	CC 94
Specified arrhythmias	CC 96
Cerebral hemorrhage	CC 99
Ischemic or unspecified stroke	CC 100
Precerebral arterial occlusion and transient cerebral ischemia	CC 101
Cerebral atherosclerosis and aneurysm	CC 102

Variable	Code(s)
Hemiplegia/hemiparesis	CC 74, 103
History of infection	CC 1, 3-7
Metastatic cancer, acute leukemia and other severe cancers	CC 8, 9
Lymphatic, head and neck, brain, and other major cancers; breast, colorectal and other major cancers	CC 10-15
Protein-calorie malnutrition	CC 21
Other significant endocrine and metabolic disorders	CC 22-26
Other gastrointestinal disorders	CC 38
Disorders of the vertebrae and spinal discs	CC 41
Osteoarthritis of hip or knee	CC 42
Other musculoskeletal and connective tissue disorders	CC 45
Iron deficiency or other unspecified anemias and blood disease	CC 49
Dementia or other specified brain disorders	CC 51-53
Major psychiatric disorders	CC 57-59
Quadriplegia, other extensive paralysis	CC 70-73
Multiple sclerosis	CC 77, 81
Seizure disorders and convulsions	CC 79
Hypertension	CC 95
Vascular disease and complications	CC 106-108
Chronic obstructive pulmonary disease (COPD)	CC 111
Pneumonia	CC 114-116
Pleural effusion/pneumothorax	CC 117
Other eye disorders	CC 128
Other ear, nose, throat, and mouth disorders	CC 131
Dialysis status	CC 134
Renal failure	CC 135-140
Urinary tract infection	CC 144
Male genital disorders	CC 149
Decubitus ulcer of skin	CC 157-160
Chronic ulcer of skin, except decubitus	CC 161
Other dermatological disorders	CC 165
National Institute of Health Stroke Scale (NIH Stroke Scale)	ICD-10 R29700-R29742

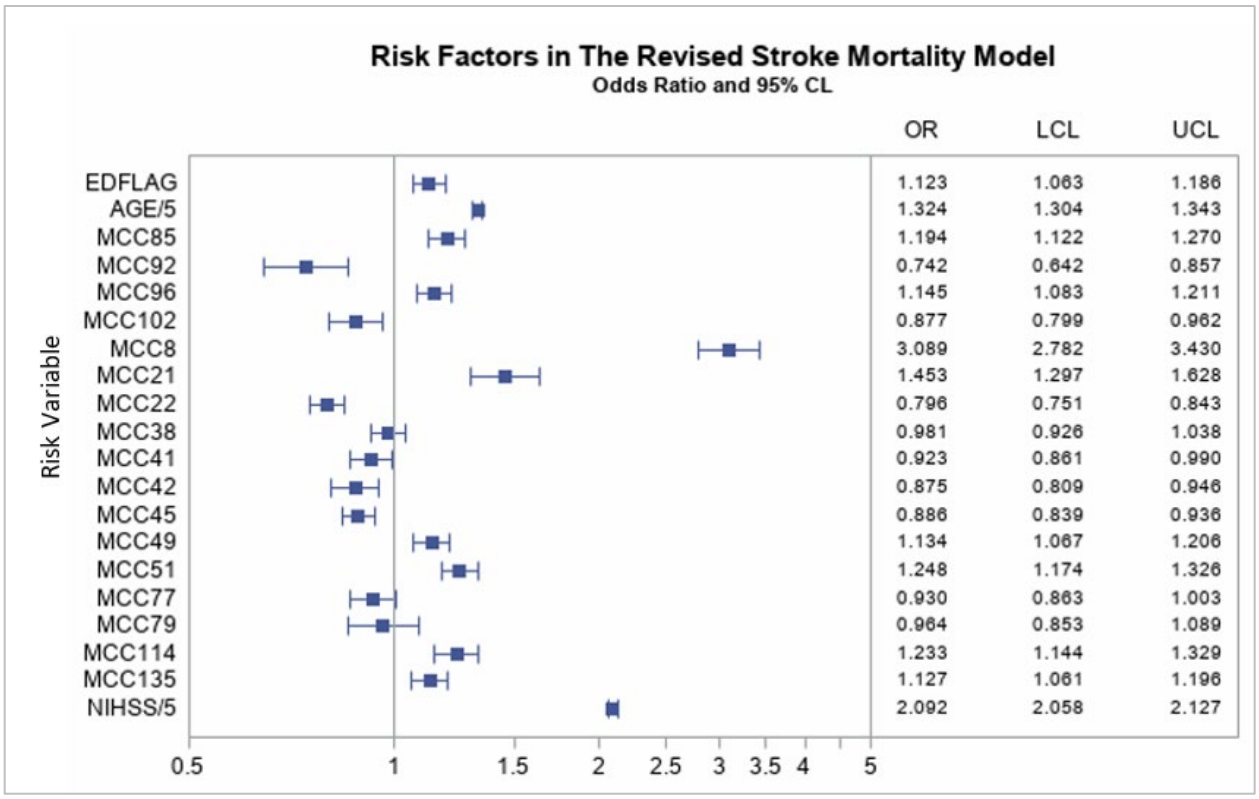
*cell intentionally left blank

Table 4. Final model variables (variables meeting criteria in field 2b4.3)

Variable	Code(s)	10/2016 – 06/2019 OR (95% CI)
Age (continuous, per 5 units)	*	1.32 (1.30, 1.34)
Transfer from another ED	*	1.12 (1.06, 1.19)
Congestive heart failure	CC 85	1.19 (1.12, 1.27)
Congenital cardiac/circulatory defects	CC 92, 93	0.74 (0.64, 0.86)
Specified heart arrhythmias	CC 96	1.15 (1.08, 1.21)
Cerebral atherosclerosis and aneurysm	CC 102	0.88 (0.80, 0.96)
Metastatic cancer and acute leukemia and other major cancers	CC 8, 9	3.09 (2.78, 3.43)
Protein - calorie malnutrition	CC 21	1.45 (1.30, 1.63)
Other significant endocrine and metabolic disorders	CC 22-26	0.80 (0.75, 0.84)
Other gastrointestinal disorders	CC 38	0.98 (0.93, 1.04)
Disorders of the vertebrae and spinal discs	CC 41	0.92 (0.86, 1.00)
Osteoarthritis of hip or knee	CC 42	0.88 (0.81, 0.95)
Other musculoskeletal and connective tissue disorders	CC 45	0.89 (0.84, 0.94)
Iron deficiency and other/unspecified anemia and blood disease	CC 49	1.13 (1.07, 1.21)
Dementia or other specified brain disorders	CC 51-53	1.25 (1.17, 1.33)
Multiple sclerosis	CC 77, 81	0.93 (0.86, 1.00)
Seizure disorders and convulsions	CC 79	0.96 (0.85, 1.09)
Pneumonia	CC 114-116	1.23 (1.14, 1.33)
Renal failure	CC 135-140	1.13 (1.06, 1.20)
National Institute of Health Stroke Scale (NIH Stroke Scale)	ICD-10 R29700-R29742	2.09 (2.06, 2.13)

*cell intentionally left blank

Figure 5. Final model risk variables and odds ratios



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

Variation in prevalence of the factor across measured entities

The prevalence of social risk factors in the Stroke Mortality with Adjustment for Stroke Severity cohort varies widely across entities. The median percentage of dual-eligible patients is 11.19% (interquartile range [IQR] 7.69%-16.85%). The median percentage of patients with a low AHRQ SES index score (lowest quartile) is 12.03% (IQR 5.26%-23.08%).

Empirical association with the outcome (univariate)

The patient-level observed stroke mortality rates are higher for dual-eligible patients (17.49%) compared with all other patients (14.14%). The mortality rate for patients with a low AHRQ SES index score was slightly higher (15.27%) compared with all other patients (14.42%).

Incremental effect of SES variables in a multivariable model

We then 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 of the claims-based clinical variables, the effect size of each of these variables is small. The dual-eligibility variable had an odds ratio of 1.29 (95% CI 1.22, 1.36) in the bivariate analysis and 0.92 (95% CI 0.86, 0.98) in the multivariate analysis. The low AHRQ SES variable had an odds ratio of 1.07 (95% CI 1.02, 1.12) in the bivariate analysis and 1.04 (95% CI 0.98, 1.11) in the multivariate analysis. Dual-eligibility was significant in the multivariable model. In all cases the c-statistics for the stroke patient-level multivariate models with the SES variables in the models were unchanged from those without (model with original variables: 0.86; model with dual-eligible variable: 0.86; model with AHRQ SES index variable: 0.86).

Table 5. Social Risk Variable Model Effects

Variable	Bivariate Analysis OR (95% CI)	Multivariate Analysis OR (95% CI)	C-Statistic
Low AHRQ SES	1.07 (1.02, 1.12)	1.04 (0.98, 1.11)	0.86
Dual-eligibility	1.29 (1.22, 1.36	0.92 (0.86, 0.98)	0.86

To further understand the relative importance of these risk-factors in the measure we compared hospital performance with and without the addition of each social risk factor variable. We find that the addition of any of these variables into the model had little to no effect on hospital performance. The mean absolute change in hospitals' RSMRs when adding a dual-eligibility indicator was 0.001% with a correlation coefficient between RSMRs for each hospital with and without dual-eligibility of 0.999. The mean absolute change in hospitals' RSMRs when adding a low SES AHRQ indicator was 0.00% with a correlation coefficient between RSMRs for each hospital with and without low SES of 0.999.

Summary

Overall, we find that among the SRF variables that could be feasibly incorporated into this model, low AHRQ SES index does not have a significant relationship with the outcome in multivariable modeling, while dual-eligible status has an association in the opposite direction than what has been the expressed concern of stakeholders interested in adding such adjustment to the models. While dual-eligibility was associated with increased mortality in the bivariate analyses, the association was reversed after accounting for comorbidities and stroke severity, indicating that any effects of dual-eligibility are already covered by clinical risk variables. We also find that the impact of any of these indicators is 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 findings and complex pathways which could explain any relationship between social risk factors and mortality, and do not all support risk-adjustment, we did not incorporate social risk factor variables in this measure.

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

Approach to assessing model performance

We computed three summary statistics for assessing model performance (Harrell and Shih, 2001) for the development and validation 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 patients from low-risk patients. 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 developed in a randomly selected 50% sample of the hospitalizations for ischemic stroke in **Dataset 1** (development dataset; October 1, 2016-June 30, 2019) by validating the model in the validation sample (the remaining 50% of the dataset).

Reference:

Harrell FE, Shih YCT. 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 sample (development sample):

- C-statistic = 0.86
- Predictive ability (lowest decile %, highest decile %) = (0.94, 60.37)

2nd half of randomly split sample (validation sample):

- C-statistic = 0.86
- Predictive ability (lowest decile %, highest decile %) = (0.95, 61.27)

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

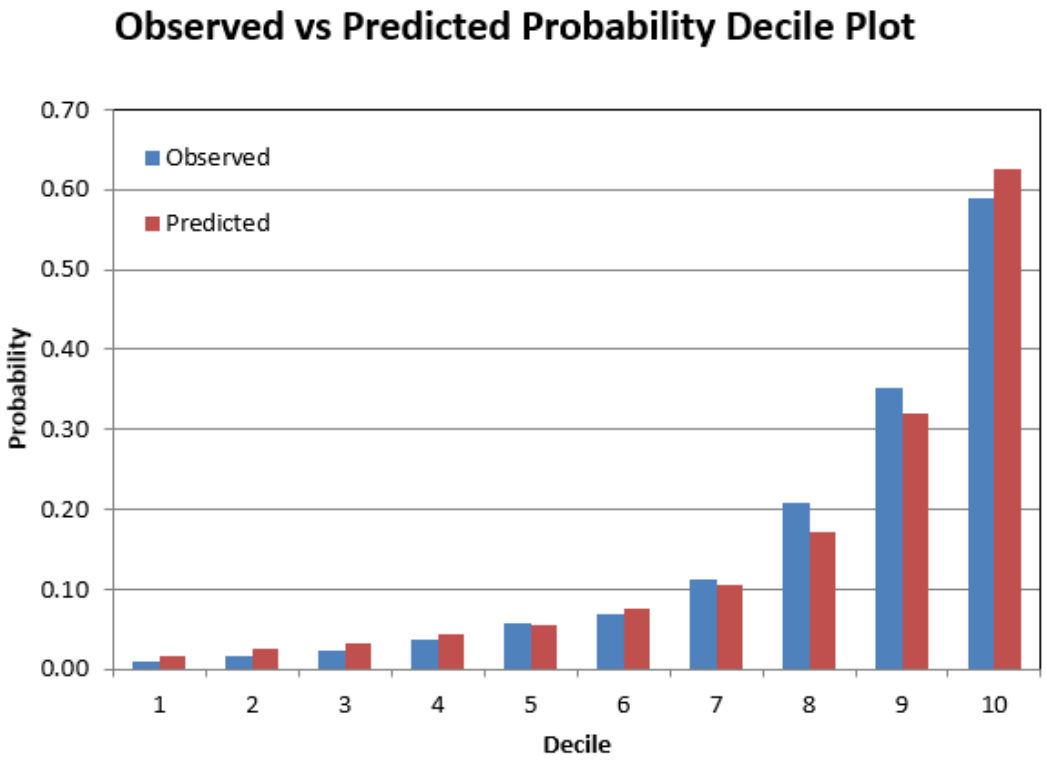
2b3.7. Statistical Risk Model Calibration Statistics (*e.g., Hosmer-Lemeshow statistic*):

Calibration in the validation sample is based on the model from the development sample (**Dataset 1**): (0.02, 1.01).

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

The probability decile plot is a graphical depiction of the observed mortality in the deciles of the predicted mortality to measure predictive ability. Below, we present the probability decile plot in the validation dataset.

Figure 6. Observed vs Predicted Probability Decile Plot (October 1, 2016-June 30, 2019)



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-statistics of 0.86 in the development sample and 0.86 in the validation sample indicate excellent model discrimination (**Dataset 1**). 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. For comparison, the Hospital 30-Day Mortality Following Acute Ischemic Stroke Hospitalization measure within IQR without adjustment for stroke severity has a c-statistic of 0.75.

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 values of almost 0 at one end and almost 1 on the other end indicate good calibration of the model.

Probability 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 excellent 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*)

N/A

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

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

The 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, we estimated the model coefficients using data from October 1, 2016 through June 30, 2019.

The method for discriminating hospital performance has not been determined. For public reporting of measures of hospital outcomes developed with similar methodology, CMS characterizes the uncertainty associated with the RSMR by estimating the 95% interval estimate. This is similar to a 95% confidence interval but is calculated differently. If the RSMR's interval estimate does not include the national observed mortality rate (is lower or higher than the rate), then CMS is confident that the hospital's RSMR is different from the national rate and describes the hospital on the Hospital Compare website as "better than the U.S. national rate" or "worse than the U.S. national rate." If the interval includes the national rate, then CMS describes the hospital's RSMR as "no different than the U.S. national rate" or "the difference is uncertain." CMS does not classify performance for hospitals that have fewer than 25 cases in the three-year period.

However, this measure is not currently publicly reported and decisions about the approach to discriminating hospital performance have not been made.

To quantify the between hospital variance, we calculated the median odds ratio (MOR). The median odds ratio represents the median increase in the odds of a readmission within 30 days of a stroke 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 (Merlo et al, 2006).

References:

Merlo, J., Chaix, B., Ohlsson, H., Beckman, A., Johnell, K., Hjerpe, P., … & 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. *Journal of Epidemiology & Community Health*, *60*(4), 290-297.

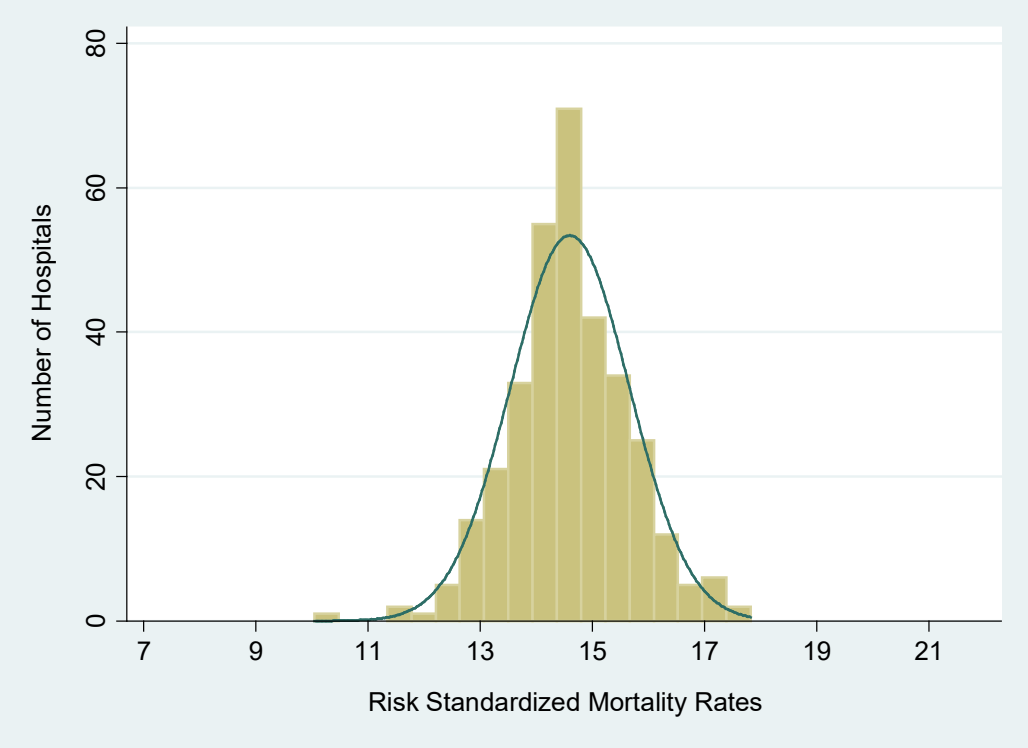
2b4.2. What were the statistical results from testing the ability to identify statistically significant and/or clinically/practically meaningful differences in performance measure scores across measured entities? (e.g., *number and percentage of entities with scores that were statistically significantly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

Analyses of Medicare FFS data show substantial variation in RSMRs among hospitals. Using data from the testing sample (October 1, 2016-June 30, 2019), the median hospital RSMR was 14.68%, with a range of 10.05% to 17.83%. The interquartile range was 13.82% - 15.52%.

Table 6. Distribution of Risk Standardized Mortality Rate (RSMR) (N=89,795)

Mean	Std. Dev.	Min	5th Percentile	10th Percentile	25th Percentile	Median	75th Percentile	90th Percentile	95th Percentile	Max
14.63	1.25	10.05	12.69	13.04	13.82	14.68	15.52	16.28	16.55	17.83

Figure 7. Distribution of Stroke Risk-Standardized Mortality Rate (RSMR) (N=329)



The median odds ratio using the between hospital variance is 1.21 and the lower and upper odds ratios are 1.16 and 1.24, respectively.

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?*)

Despite recent decreases in mortality rates nationally, stroke is the fifth most common cause of death in the United States, affecting approximately 795,000 people annually, and has a 30-day mortality rate that varies by age from 9% in patients 65 to 74 years of age, 13.1% in those 74 to 84 years of age, and 23% in those ≥85 years of age (Virani et al., 2014; Murphy et al., 2017; Casper et al., 2008).

The variation in RSMRs suggests that there are differences in the quality of care received across hospitals for stroke that support measurement to reduce this variation.

The median odds ratio using the between hospital variance demonstrates variability in hospital performance. The median odds ratio suggests a meaningful increase in the risk of readmission if a patient is admitted with stroke at a higher risk hospital compared to a lower risk hospital. A value of 1.21 indicates that a patient's risk of readmission is 21% times greater in a higher risk hospital than a lower risk hospital.

References:

Casper ML, Nwaise IA, Croft JB, Nilasena DS. Atlas of Stroke Hospitalizations Among Medicare Beneficiaries. Atlanta, GA: US Department of Health and Human Services, Centers for Disease Control and Prevention; 2008.

Virani SS, Alonso A, Benjamin EJ, et al. Heart Disease and Stroke Statistics-2020 Update: A Report From the American Heart Association. Circulation. 2020;141(9):e139 - e596. doi:10.1161/CIR.0000000000000757

Murphy SL, Xu J, Kochanek KD, Arias E. Mortality in the United States, 2017. NCHS Data Brief. 2018;(328):1 - 8.

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*)

This measure is intended to improve upon the existing stroke mortality measure currently reported within IQR by including risk adjustment for stroke severity upon admission. This measure would utilize the same outcome, cohort, and 3-year measurement timeframe. Hospital reporting of the NIH Stroke Scale is low but continuously increasing. Eventually, we anticipate the measure will be implemented with multiple imputation to address missing NIH Stroke Scale scores. While most of the literature on multiple imputation focuses on the type of missing data, rather than the proportion of missing data, (Dong et al, 2013; Enders et al, 2017; JAC et al, 2009; Madley-Dowd et al, 2019) external statistical experts advised that multiple imputation should only be used when at least 60% of data are available.

Upon implementation, if hospital reporting of the NIH Stroke Scale for ischemic stroke admissions is still too low, the measure could alternatively use simple replacement with zero to address missing NIH Stoke Scale scores. This would further align with the guidelines put forth by AHA/ASA by incentivizing hospital reporting of the NIH Stroke Scale.

To test the impact that using simple replacement with zero, instead of multiple imputation, we 1) produced a model c-statistic using simple replacement with zero on all hospitals and 2) compared hospital performance by quintiles of RSMRs between the Hospital 30-Day Mortality Following Acute Ischemic Stroke Hospitalization measure currently reported in IQR without risk adjustment for stroke severity and the revised measure with risk adjustment for stroke severity and simple replacement with zero to address missing NIH Stoke Scale scores.

References:

Dong, Y., & Peng, C. Y. J. (2013). Principled missing data methods for researchers. *SpringerPlus*, *2*(1), 222.

Enders, C. K. (2017). Multiple imputation as a flexible tool for missing data handling in clinical research. *Behaviour research and therapy*, *98*, 4-18.

JAC, S. (2009). White IR., Carlin JB., Spratt M., Royston P., Kenward MG., Wood AM., Carpenter JR. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ*, *338*, b2393.

Madley-Dowd, P., Hughes, R., Tilling, K., & Heron, J. (2019). The proportion of missing data should not be used to guide decisions on multiple imputation. *Journal of clinical epidemiology*, *110*, 63-73.

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*)

The measure with simple replacement with zero to address missing NIH Stoke Scale scores produced a model c-statistic of 0.76.

Below is a table demonstrating the impact of risk adjustment for stroke severity and simple replacement with zero on hospital performance by providing a cross tab of RSMR quintiles for the Hospital 30-Day Mortality Following Acute Ischemic Stroke Hospitalization measure in IQR and the revised measure with risk adjustment for stroke severity and simple replacement with zero.

Table 7. Cross tabulation of number of hospitals by RSMR quintiles in the Hospital 30-Day Mortality Following Acute Ischemic Stroke Hospitalization measure in IQR and the revised Stroke Mortality with Adjustment for Stroke Severity and simple replacement with zero

Stroke Mortality in IQR	Stroke Mortality with Adjustment for Stroke Severity and Simple Replacement with Zero: Q1	Stroke Mortality with Adjustment for Stroke Severity and Simple Replacement with Zero: Q2	Stroke Mortality with Adjustment for Stroke Severity and Simple Replacement with Zero: Q3	Stroke Mortality with Adjustment for Stroke Severity and Simple Replacement with Zero: Q4	Stroke Mortality with Adjustment for Stroke Severity and Simple Replacement with Zero: Q5	Total
Q1	667	152	16	7	0	842
Q2	81	530	174	48	10	843
Q3	23	91	550	162	17	843
Q4	37	36	63	529	178	843
Q5	34	34	40	97	638	843
Total	842	843	843	843	843	4,214

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*)

The measure model performance improves slightly with risk adjustment for stroke severity and simple replacement with zero (0.76), compared to the stoke mortality measure current reported in IQR (0.75).

When using simple replacement with zero to address missing NIH Stroke Scale scores, most hospitals will remain in the same quintile or move to a neighboring quintile (92.8%) from the stroke mortality without risk adjustment for stroke severity currently reported within IQR. This demonstrates limited impact of using simple replacement with zero on hospital RSMRs.

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*)

To evaluate whether performance of the risk measure is not overly biased by missing stroke severity data we analyzed the risk measure using multiple imputation for missing values. Based on the advice from external statistical experts, to demonstrate validity of the measure, we limited the testing sample to hospitals that report the NIH Stroke Scale for at least 60% of ischemic stroke admissions and used multiple imputation for missing values.

The multiple imputation technique used to impute missing values was a multi-logit regression model. Five copies of imputation datasets were produced for the analyses. The results based on these data were aggregated according to the standard statistical methods for the presentation of the results and for the measure score calculation.

In multiple imputation, missing variable values are predicted using other related patient variables available. The predicted values are substituted for the missing values, which results in a full data set without any missing variables (the imputed data set). By repeating this process multiple times, we get multiple imputed data sets. We then conduct analyses on and obtain results for each imputed data set. The results based on multiple data sets are combined to produce the overall final results. The multiple imputation represents a random sample of the missing values according to the association of the non-missing values of all the variables considered, and the resulting inferences of multiple imputation are statistically valid, which reflect the uncertainty due to missing values (Van Buuren, 2007).

To ensure the imputation of NIH Stroke Scale using multiple imputation was valid, we compared the distribution of the NIH Stroke Scale among the admissions with reported NIH Stroke Scale scores and the distribution of NIH Stroke Scale for all admissions after imputation, and compared their association with 30-day mortality.

We examined characteristics of hospital reporting of the NIH Stroke Scale by hospital characteristics in our testing sample of hospitals that report the NIH Stroke Scale for at least 60% of ischemic stroke admissions from October 1, 2016 to June 30, 2019. We also examined hospital reporting of the NIH Stroke Scale by hospital characteristics in the most recent full year of available data (July 2018 – June 2019) to demonstrate the increased reporting of the NIH Stroke Scale.

References:

He R, Belin T. Multiple imputation for high-dimensional mixed incomplete continuous and binary data. Stat. Med. 2014;33:2251–2262.

Dong, Y., & Peng, C. Y. J. (2013). Principled missing data methods for researchers. *SpringerPlus*, *2*(1), 222.

Enders, C. K. (2017). Multiple imputation as a flexible tool for missing data handling in clinical research. *Behaviour research and therapy*, *98*, 4-18.

JAC, S. (2009). White IR. Carlin JB. Spratt M. Royston P. Kenward MG. Wood AM. Carpenter JR. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ*, *338*, b2393.

Madley-Dowd, P., Hughes, R., Tilling, K., & Heron, J. (2019). The proportion of missing data should not be used to guide decisions on multiple imputation. *Journal of clinical epidemiology*, *110*, 63-73.

Van Buuren, S. (2007), "Multiple Imputation of Discrete and Continuous Data by Fully Conditional Specification," *Statistical Methods in Medical Research*, 16, 219–242.

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*)

While most of the literature on multiple imputation focuses on the type of missing data, rather than the proportion of missing data, (Dong et al, 2013; Enders et al, 2017; JAC et al, 2009; Madley-Dowd et al, 2019) external statistical experts advised that multiple imputation should only be used when at least 60% of data are available. Therefore, the testing sample was based on hospitals that report the NIH Stroke Scale for at least

60% of ischemic stroke admissions included 329 hospitals and 89,795 admissions. In this testing sample, NIH Stroke Scale scores were available in 71.71% of patients with an admission for ischemic stroke from October 1, 2016 to June 30, 2019.

We used multiple imputation to address missing NIH Stroke Scale scores. Below is a graph that plots the NIH Stroke Scale scores and 30-day mortality rate without (blue line) and with imputation (green line).

Figure 8. Mortality rates by NIH Stroke Scale scores before and after multiple imputation

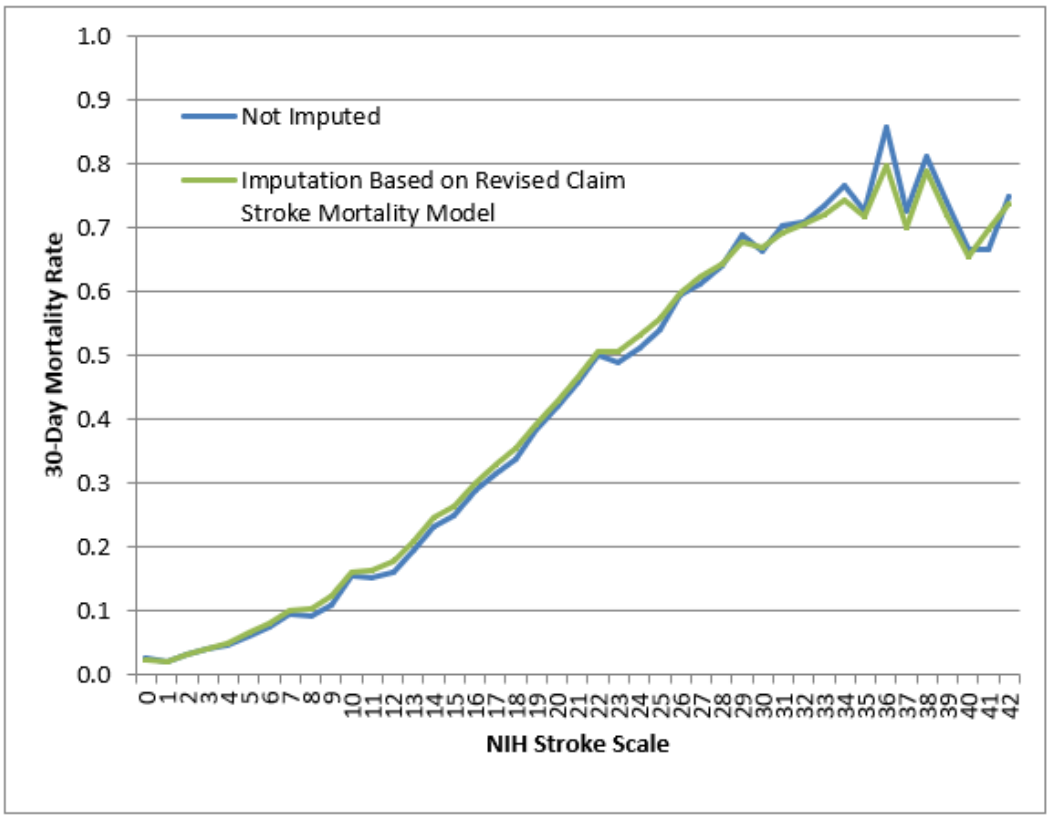


Table 8. Hospital reporting of the NIH Stroke Scale by hospital characteristics within testing sample (October 1, 2016 – June 30, 2019)

Description	Total: #	Total: Column %	Hospitals reporting NIH Stroke Scale ≥60%: #	Hospitals reporting NIH Stroke Scale ≥60%: Row %
All	4,214	100.00	329	100.00
Number of beds: Missing*	96	2.28	6	1.82
Number of beds: < 300	3,366	79.88	163	49.54
Number of beds: 300 to 600	571	13.55	94	28.57
Number of beds: > 600	181	4.30	66	20.06

Description	Total: #	Total: Column %	Hospitals reporting NIH Stroke Scale ≥60%: #	Hospitals reporting NIH Stroke Scale ≥60%: Row %
Teaching status: Missing*	96	2.28	6	1.82
Teaching status: COTH	227	5.39	92	27.96
Teaching status: Teaching	1,092	25.91	109	33.13
Teaching status: Non-Teaching	2,799	66.42	122	37.08
Core based statistical area: Missing*	96	2.28	6	1.82
Core based statistical area: Metro	2,463	58.45	286	86.93
Core based statistical area: Micro	717	17.01	21	6.38
Core based statistical area: Rural	938	22.26	16	4.86
Safety Net Hospital: Missing*	97	2.30	6	1.82
Safety Net Hospital: No	2,933	69.60	259	78.72
Safety Net Hospital: Yes	1184	28.10	64	19.45
Critical Access Hospital: Missing*	96	2.28	6	1.82
Critical Access Hospital: No	3,007	71.36	309	93.92

Description	Total: #	Total: Column %	Hospitals reporting NIH Stroke Scale ≥60%: #	Hospitals reporting NIH Stroke Scale ≥60%: Row %
Critical Access Hospital: Yes	1,111	26.36	14	4.26

*Some hospitals could not be linked to the AHA data

Table 9. Hospital reporting of the NIH Stroke Scale by hospital characteristics (July 2018 – June 2019)

Description	Total: #	Total: Column %	Hospitals reporting NIHSS >60%: #	Hospitals reporting NIHSS >60%: Row %
All	3,823	100.00	924	100.00
Number of beds: Missing*	79	2.07	19	2.06
Number of beds: < 300	2,995	78.34	554	59.96
Number of beds: 300 to 600	568	14.86	239	25.87
Number of beds: > 600	181	4.73	112	12.12
Teaching status: Missing*	79	2.07	19	2.06
Teaching status: COTH	222	5.81	135	14.61
Teaching status: Teaching	1,067	27.91	344	37.23
Teaching status: Non-Teaching	2,455	64.22	426	46.10
Core based statistical area: Missing*	79	2.07	19	2.06
Core based statistical area: Metro	2,325	60.82	769	83.23
Core based statistical area: Micro	669	17.50	87	9.42

Description	Total: #	Total: Column %	Hospitals reporting NIHSS >60%: #	Hospitals reporting NIHSS >60%: Row %
Core based statistical area: Rural	750	19.62	49	5.30
Safety Net Hospital: Missing*	80	2.09	19	2.06
Safety Net Hospital: No	2,720	71.15	760	82.25
Safety Net Hospital: Yes	1,023	26.76	145	15.69
Critical Access Hospital: Missing*	79	2.07	19	2.06
Critical Access Hospital: No	2,874	75.18	857	92.75
Critical Access Hospital: Yes	870	22.76	48	5.19

*Some hospitals could not be linked to the AHA data

References:

Dong, Y., & Peng, C. Y. J. (2013). Principled missing data methods for researchers. *SpringerPlus*, *2*(1), 222.

Enders, C. K. (2017). Multiple imputation as a flexible tool for missing data handling in clinical research. *Behaviour research and therapy*, *98*, 4-18.

JAC, S. (2009)., White IR., Carlin JB., Spratt M., Royston P., Kenward MG., Wood AM., Carpenter JR. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ*, *338*, b2393.

Madley-Dowd, P., Hughes, R., Tilling, K., & Heron, J. (2019). The proportion of missing data should not be used to guide decisions on multiple imputation. *Journal of clinical epidemiology*, *110*, 63-73.

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*)

Hospital reporting of the NIH Stroke Scale was relatively high (71.17%) within the testing sample.

When comparing the mortality rates of the admissions with NIH Stroke Scale scores with all admissions after multiple imputation, as seen in Figure 8, the mortality rates are very similar, demonstrating that the multiple imputation approach is valid across the range of NIH Stroke Scale scores.

Overall, most hospital characteristics had at least some hospitals reporting the NIH Stroke Scale for at least 60% of ischemic stroke admissions and therefore were represented within our testing sample. There were fewer small, non-teaching, suburban, and rural hospitals, as well as safety-net hospitals and critical access hospitals. However, although the testing sample was not representative of the distribution of all hospital

characteristics, we do have representation of these characteristics within the testing sample and many of these characteristics are also found in the overall population of hospitals.

When examining the most recent full year of available data, the number of hospitals reporting the NIH Stroke Scale for at least 60% of ischemic stroke admissions increased substantially from 329 on October 1, 2016 - June 30, 2019, to 924 in July 2018 - June 2019; this increase in reporting can be observed across all hospital characteristics. This demonstrates that hospital reporting of the NIH Stroke Scale is continuously increasing across all hospital characteristics. While we had to test the measure using 33 months of historical data with lower reporting of the NIH Stroke Scale, to present testing that aligns with the intended 3-year timeframe of the measure, the measure would be implemented with newer data and substantially more NIH Stroke Scale data present.

3. Feasibility

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

3a. Byproduct of Care Processes

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

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

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

If other:

3b. Electronic Sources

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

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

ALL data elements are in defined fields in a combination of electronic sources

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

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

Attachment:

3c. Data Collection Strategy

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

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

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

Currently there is no national database that includes NIH Stroke Scale score data for stroke patients admitted to all non-federal acute care hospitals. Therefore, implementation of this measure depends on hospitals including patients' NIH Stroke Scale scores, for all patients admitted with acute ischemic stroke, in the claims they submit to Medicare using ICD-10 codes. Collection of the NIH Stroke Scale is now Class I recommended in the AHA/ASA guidelines for care of patients admitted with acute ischemic stroke. New ICD-10 codes for NIH Stroke Scale scores became available to hospitals to include in Medicare claims, which are routinely collected as part of the billing process, in October 2016.

Based on all acute care hospitals from October 1, 2016 to June 30, 2019, NIH Stroke Severity scores were available in 37.03% of patients with an admission for ischemic stroke. This proportion of admissions with NIH Stroke Scale scores increased from 13.33% in October 2016 to 55.59% in May 2019. These rates demonstrate a low but gradually increasing rate of reporting and CMS expects rates of NIH Stroke Scale reporting to continue to increase.

For measure testing, we limited our sample to acute care hospitals that report the NIH Stroke Scale for at least 60% of ischemic stroke admissions with multiple imputations to address missing NIH Stroke Scale scores. It is in the intention for this measure to use multiple imputation, as demonstrated to be a valid approach within this submission once hospitals report a sufficient threshold of NIH Stroke Scale scores. While most of the literature on multiple imputation focuses on the type of missing data, rather than the proportion of missing data (Dong et al, 2013; Enders et al, 2017; JAC et al, 2009; Madley-Dowd et al, 2019), external statistical experts advised that multiple imputation should only be used when at least 60% of data are available. Therefore, CMS announced that they will use simple replacement with zero to address missing data during confidential reporting and initial implementation to address missing NIH Stroke Scale scores and further incentivize NIH Stroke Scale reporting. CMS will evaluate this strategy as the reporting of the NIH Stroke Scale increases over time.

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 Not in use	*

*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

4a1.2. If not currently publicly reported OR used in at least one other accountability application (e.g., payment program, certification, licensing) what are the reasons? (*e.g., Do policies or actions of the developer/steward or accountable entities restrict access to performance results or impede implementation?*)
This is a new measure with a plan for use, outlined below in 4a1.3

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.*)

CMS intends to implement this measure in the Hospital Inpatient Quality Reporting (Hospital IQR) Program to replace the currently reported Hospital 30-Day Mortality Following Acute Ischemic Stroke Hospitalization measure.

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

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

This measure is not yet implemented.

However, it is a re-specified version of the currently reported Hospital 30-Day Mortality Following Acute Ischemic Stroke Hospitalization measure that is implemented in the Inpatient Quality Reporting Program. The measure was re-specified based on stakeholder feedback that the measure should risk adjust for stroke severity upon admission. In 2016, a workgroup consisting of neurologists, cardiologists and experts in biostatistics, measurement, and quality improvement was convened to provide clinical expertise on the measure. The work group met regularly throughout development to address key issues related to measure cohort, outcome, and usability. In addition, we also posted the measure specifications for public comment, which resulted in overall agreement on the inclusion of the NIH Stroke Scale as a risk adjustment.

For the currently reported measure, the exact number of measured entities (acute care hospitals) varies with each new measurement period. In 2020, 4,254 hospitals were included in the measure calculation. These were all non-federal short-term acute care hospitals (including Indian Health Service hospitals) and critical access hospitals with at least 25 stroke admissions between July 2016 and June 2019. In 2017, the measure calculation included admissions from 4,417 hospitals. While the number of measured entities may vary slightly from year to year, the measured entities are the same for the re-specified measure.

The measured entities (hospitals that provide acute inpatient and outpatient care) 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, and in the re-specification of this measure.

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.

Similarly to all of CMS's publicly reported mortality and readmission measures, including the current stroke mortality measure, each hospital receives confidential measure results prior to annual public reporting each calendar year through CMS's QualityNet website. The results are then publicly reported on CMS's Care Compare website each calendar year.

While the code used to process the claims data and calculate measure results is written in SAS and available upon request, the measure is risk standardized using data from all hospitals and thus, stakeholders cannot independently calculate measure scores. 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 (HSR) 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 patients transferred in from particular facilities).

The HSRs 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.

Each year, hospitals have access to the following list of updated resources provided directly or posted publicly for use:

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

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

Similar resources and processes will be provided and maintained for this re-specified version of the measure.

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)

For the measure currently in public reporting, the measured entities (hospitals that provide acute inpatient and outpatient care) 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, and in the re-specification of this measure.

Literature Reviews

In addition, we routinely scan literature repositories 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.

The majority of inquiries received from hospitals through the Q&A process relate to clarifying questions about data sources and aspects of the methodology, specific questions related to hospital performance outlined within Hospital-Specific Reports, and requests for code set files and SAS code.

However, since CMS signaled in the FY 2017 IPPS Final rule the refinement of the Hospital 30-Day Mortality Following Acute Ischemic Stroke Hospitalization measure to include the NIH Stroke Scale for risk-adjustment, we received the following feedback:

1. Comments emphasizing the need to adjust for stroke severity upon admission using NIH Stroke Scale ICD-10 codes; and
2. Requests for detailed measure specifications regarding how CMS intends to use the NIH Stroke Scale in risk adjustment.

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

The majority of inquiries received from researchers, patients, and other individuals through the Q&A process relate to clarifying questions about data sources and aspects of the methodology as well as requests for code set files and SAS code.

We have also received requests for detailed measure specifications including reporting the NIH Stroke Scale and its use in risk adjustment.

Additionally, stakeholder feedback was obtained through the original Technical Expert Panel (2010) and the 2012 NQF Neurology Steering Committee. In 2016, a workgroup consisting of neurologists, cardiologists and experts in biostatistics, measurement, and quality improvement was convened to provide clinical expertise and feedback on the measure. Through the FY2018 IPPS rulemaking public comment process, CMS received stakeholder feedback urging that the measure currently in public reporting risk adjust for stroke severity upon admission using the NIH Stroke Scale.

Summary of Relevant Publications from the Literature Review:

Since 2016, we have reviewed more than 250 articles related to mortality following hospitalization for acute ischemic stroke. Relevant articles shared key themes related to: considerations for additional risk adjustment variables, including social risk factors and other clinical comorbidities; national trends in stroke mortality and

geographic variation; racial disparities in stroke mortality rates; comparison of stroke mortality rates in primary stroke centers (PSCs) compared to non-PSCs; and, examination of NIH Stroke Scale validation and impact on stroke mortality measure model performance.

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.

Of note, we received stakeholder feedback that the currently reported stroke mortality measure should adjust for stroke severity upon admission using the NIH Stroke Scale. As a result, in 2016 we re-specified the measure to include the NIH Stroke Scale within the risk adjustment model and submitted the measure to NQF for initial endorsement. However, because we could only use registry-based stroke scale scores for testing, the measure was not NQF endorsed at that time. Now, in 2020, we present the re-specified measure, using the NIH Stroke Scale to adjust for stroke severity, with measure testing results using 33 months of available ICD-10 data.

Each year, issues raised through the Q&A process or in the literature related to the currently reported stroke mortality 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 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 receives public comment on the changes and finalizes those changes in the IPPS or another rule. A similar process will be maintained for this re-specified version of the measure.

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 RSMR in the 2013-2016 dataset was 14.5%. This dataset included 519,732 admissions from 4,417 hospitals. The median hospital RSMR in the 2016-2019 combined dataset was 13.6% based on 520,432 admissions from 4,254 hospitals. This decline suggests that there is opportunity for further improvement in the 30-day mortality outcome over time. Once this measure is ready for public reporting, CMS intends to replace the currently publicly reported stroke measure that does not adjust for severity. This new stroke measure has improved credibility and face validity among stakeholders and increased the ability to differentiate hospital performance using the NIH Stroke Scale.

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.

We did not identify any unexpected benefits during measure development or model testing.

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)

0467 : Acute Stroke Mortality Rate (IQI 17)

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.

Competing Measure

Hospital 30-Day, All-Cause, Risk-Standardized Mortality Rate Following Acute Ischemic Stroke (not NQF endorsed)

5a. Harmonization of Related Measures

The measure specifications are harmonized with related measures;

OR

The differences in specifications are justified

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

Are the measure specifications harmonized to the extent possible?

Yes

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

This measure and the NQF endorsed Acute Stroke Mortality Rate (IQI 17) (AHRQ) Measure #0467 are complementary and related rather than competing measures. Although they both assess mortality for patients admitted to acute care hospitals with a principal discharge diagnosis of acute ischemic stroke, the specified outcomes are different. Our measure assesses 30-day mortality, while #0467 assesses inpatient mortality. The 30-day mortality and inpatient mortality outcomes each have distinct advantages and uses, which make them complementary (and related) as opposed to competing. For example, the 30-day period provides a broader perspective on hospital care and utilizes a standard time period to examine hospital performance to avoid bias by differences in length of stay among hospitals. However, in some settings it may not be feasible to capture post-discharge mortality, making the inpatient measure more useable. We have previously consulted with AHRQ to examine harmonization of the measures' cohort. Because of that collaboration, we have found that the measures' cohorts are harmonized to the extent possible and that the small differences in cohort inclusion and exclusion criteria are appropriate because the measures assess different outcomes. We did not include in our list of related measures any non-outcome (such as process) measures with the same target population as our measure. 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.)

This measure looks at a longer outcome time frame (30-days versus in-hospital) than the NQF endorsed Acute Stroke Mortality Rate (IQI 17) (AHRQ) Measure #0467 and incorporates stroke severity into the risk-model. The current publicly reported measure, Hospital 30-Day Mortality Following Acute Ischemic Stroke Hospitalization Measure, is not a potentially competing measure since it is CMS' intent to replace the current measure with this newly re-specified measure, which includes stroke severity in the risk model.

Appendix

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

Attachment:

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 (YNHHSC/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.

Our working group consisted of the following members:

- Lee Schwamm, MD: Vice Chairman, Department of Neurology, Massachusetts General Hospital. Provided experience relevant to clinical content and/or performance measurement.

- Gregg Fonarow, MD: Professor of Medicine, University of California, Los Angeles. Provided experience relevant to clinical content and/or performance measurement.
- Jason Sico, MD: Director, Stroke Care VA Connecticut Healthcare System. Provided experience relevant to clinical content and/or performance measurement.
- Kevin Sheth, MD: Associate Professor of Neurology and Neurosurgery, Yale University. Provided experience relevant to clinical content and/or performance measurement.

Measure Developer/Steward Updates and Ongoing Maintenance

Ad.2 Year the measure was first released:

Ad.3 Month and Year of most recent revision:

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

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

Ad.6 Copyright statement: N/A

Ad.7 Disclaimers: N/A

Ad.8 Additional Information/Comments: N/A