NATIONAL QUALITY FORUM

MEASURE WORKSHEET

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

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

Purple text represents the responses from measure developers.

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

Brief Measure Information

NQF #: 3032

Corresponding Measures:

De.2. Measure Title: STS Mitral Valve Repair/Replacement (MVRR) + Coronary Artery Bypass Graft (CABG) Composite Score

Co.1.1. Measure Steward: The Society of Thoracic Surgeons

De.3. Brief Description of Measure: The STS Mitral Valve Repair/Replacement (MVRR) + Coronary Artery Bypass Graft (CABG) Composite Score measures surgical performance for MVRR + CABG with or without concomitant Atrial Septal Defect (ASD) and Patient Foramen Ovale (PFO) closures, tricuspid valve repair (TVr), or surgical ablation for atrial fibrillation (AF). To assess overall quality, the STS MVRR +CABG Composite Score comprises two domains consisting of six measures:

Domain 1 – Absence of Operative Mortality

Proportion of patients (risk-adjusted) who do not experience operative mortality. Operative mortality is defined as death before hospital discharge or within 30 days of the operation.

Domain 2 – Absence of Major Morbidity

Proportion of patients (risk-adjusted) who do not experience any major morbidity. Major morbidity is defined as the occurrence of any one or more of the following major complications:

1. Prolonged ventilation,
2. Deep sternal wound infection,
3. Permanent stroke,
4. Renal failure, and
5. Reoperations for bleeding, coronary graft occlusion, prosthetic or native valve dysfunction, and other cardiac reasons, but not for other non-cardiac reasons.

Outcome data are collected on all patients and from all participants. For optimal measure reliability, participants meeting a volume threshold of at least 25 cases over 3 years receive a score for each of the two domains, plus an overall composite score. The overall composite score is created by "rolling up" the domain scores into a single number. In addition to receiving a numeric score, participants are assigned to rating categories designated by the following:

1 star – lower-than-expected performance

2 stars – as-expected performance

3 stars – higher-than-expected performance

1b.1. Developer Rationale: N/A

S.4. Numerator Statement: Due to the complex methodology used to construct the composite measure, it is impractical to separately discuss the numerator and denominator. The following discussion describes how each domain score is calculated and how these are combined into an overall composite score.

The STS Mitral Valve Repair/Replacement (MVRR) Composite Score comprises two domains consisting of six measures:

Domain 1 – Absence of Operative Mortality

Proportion of patients (risk-adjusted) who do not experience operative mortality. Operative mortality is defined as death before hospital discharge or within 30 days of the operation.

Domain 2 – Absence of Major Morbidity

Proportion of patients (risk-adjusted) who do not experience any major morbidity. Major morbidity is defined as the occurrence of any one or more of the following major complications:

1. Prolonged ventilation,
2. Deep sternal wound infection,
3. Permanent stroke,
4. Renal failure, and
5. Reoperations for bleeding, coronary graft occlusion, prosthetic or native valve dysfunction, and other cardiac reasons, but not for other non-cardiac reasons.

Participants receive a score for each of the two domains, plus an overall composite score. The overall composite score was created by "rolling up" the domain scores into a single number. In addition to receiving a numeric score, participants are assigned to rating categories designated by the following:

1 star – lower-than-expected performance

2 stars – as-expected performance

3 stars – higher-than-expected performance

Patient Population: The analysis population consists of patients aged 18 years or older who MVRR + CABG with or without concomitant Atrial Septal Defect (ASD) and Patient Foramen Ovale (PFO) closures, tricuspid valve repair (TVr), or surgical ablation for atrial fibrillation (AF).

Time Window: 3 years

Data Completeness Requirement: Participants are excluded from the analysis if they have fewer than 25 MVRR + CABG procedures in the patient population.

Estimation of Composite Scores and Star Ratings:

To be consistent with the conventions of previous composite measures, risk-adjusted event rates were first converted into risk-adjusted absence-of-event rates. To calculate the composite, participant-specific absence of mortality rates and absence of morbidity rates were weighted inversely by their respective standard deviations across participants. This procedure was equivalent to first rescaling the absence of mortality rates and absence of morbidity rates by their respective standard deviations across participants, and then assigning equal weighting to the rescaled rates. Finally, in order to draw statistical inferences about participant performance, a Bayesian credible interval surrounding each participant's composite score was calculated. Unlike frequentist confidence intervals, Bayesian credible intervals have an intuitively direct interpretation as an interval containing the true value of the composite score with a specified probability (e.g., 95%). To determine star ratings for each participant, the credible interval of its composite score was compared with the STS average. Participants whose intervals were entirely above the STS average were classified as 3-star (higher

than expected performance), and participants whose intervals were entirely below the STS average were classified as 1-star (lower than expected performance). Credible intervals based on different probability levels (90%, 95%, 98%) were explored, and the resulting percentages of 1, 2, and 3-star programs were calculated.

S.6. Denominator Statement: See response in S.4. Numerator Statement for complete description of measure specifications.

Patient Population: The analysis population consists of patients aged 18 years or older who MVRR + CABG with or without concomitant Atrial Septal Defect (ASD) and Patient Foramen Ovale (PFO) closures, tricuspid valve repair (TVr), or surgical ablation for atrial fibrillation (AF).See response in S.4. Numerator Statement for complete description of measure specifications.

Patient Population: The analysis population consists of patients aged 18 years or older who MVRR + CABG with or without concomitant Atrial Septal Defect (ASD) and Patient Foramen Ovale (PFO) closures, tricuspid valve repair (TVr), or surgical ablation for atrial fibrillation (AF).

S.8. Denominator Exclusions: Data Completeness Requirement: Participants are excluded from the analysis if they have fewer than 25 MVRR + CABG procedures in the patient population.

De.1. Measure Type: Composite

S.17. Data Source: Registry Data

S.20. Level of Analysis: Clinician : Group/Practice, Facility

IF Endorsement Maintenance – Original Endorsement Date: Jan 25, 2017 **Most Recent Endorsement Date:** Jan 25, 2017

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: Maintenance of Endorsement

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

Criteria 1: Importance to Measure and Report

1a. Evidence

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

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

Summary of prior review in 2016

- This composite measure encompasses patients undergoing combined coronary bypass and mitral valve surgery. It comprises two domains: risk-adjusted operative mortality and risk-adjustment major morbidity. Operative mortality includes death before hospital discharge or within 30 days of the operation. Major morbidity includes prolonged ventilation, deep sternal wound infection, permanent stroke, renal failure, and reoperation for cardiac reasons.
- The components of this composite are outcomes for which the required evidence is identification of a relationship between the outcome and at least one healthcare action that could achieve change in measure results. The developers provided information regarding service and/or care to affect mortality and each of the five morbidities.
- The developer provided references that address operative mortality and morbidity dating from the 1990s through 2014, including those related to current STS adult cardiac surgery risk models.

Changes to evidence from last review

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

☐ The developer provided updated evidence for this measure:

Updates:

Question for the Committee:

- *Is there at least one thing that the provider can do to achieve a change in the measure results?*
- *The developer attests the underlying evidence for the measure has not changed since the last NQF endorsement review. Does the Committee agree the evidence basis for the measure has not changed and there is no need for repeat discussion and vote on Evidence?*

Guidance from the Evidence Algorithm

Assess performance on outcome (Box 1) à Relationship between outcome and healthcare action (Box 2) à 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.

- Distribution of STS mitral valve repair/replacement (MVRR) + CABG measure from two consecutive time periods, January 2016 – December 2018 and January 2017 – December 2019 for participants with at least 25 eligible cases

Measures	Jan 2016 – Dec 2018	Jan 2017 – Dec 2019
# of Participants	289	272
# Operations	16,175	15,087
Mean	0.866	0.864
STD	0.02745	0.02595
IQR	0.352	0.328
0%	0.741	0.768
10%	0.831	0.831
20%	0.845	0.844
30%	0.854	0.854
40%	0.863	0.861
50%	0.869	0.866
60%	0.875	0.871

Measures	Jan 2016 – Dec 2018	Jan 2017 – Dec 2019
70%	0.882	0.878
80%	0.889	0.885
90%	0.897	0.894
100%	0.936	0.921

Disparities

- Disparities data is presented by domain for insurance status, race, and ethnicity.

Risk-adjusted odds ratios

Measures	Mortality Adjusted Odd Ratio (95% CI)	p-value	Major Morbidity Adjusted Odd Ratio (95%CI)	p-value
Insurance status among patients age >= 65	*	*	*	*
Medicare without Medicaid/Commercial-HNO	Ref	*	Ref	*
Medicare + Medicaid dual eligible	0.94 (0.71, 1.24)	0.6578	0.81 (0.68, 0.98)	0.0287
Medicare + Commercial-HMO without Medicaid	0.97 (0.84, 1.13)	0.7131	0.98 (0.90, 1.07)	0.6597
Commercial-HMO without Medicare	0.84 (.064, 1.09)	0.1880	1.04 (0.88, 1.22)	0.6680
Insurance status among patients age < 65	*	*	*	*
Commercial-HMO without Medicare/Medicaid	Ref	*	Ref	*
Medicare or Medicaid	1.17 (0.96, 1.42)	0.1265	1.09 (0.98, 1.22)	0.1148
None/Self Paid	0.97 (0.65, 1.45)	0.8796	1.02 (0.83, 1.25)	0.8393
Other	1.23 (0.77, 1.97)	0.3833	1.00 (0.76, 1.31)	0.9743
Black race	0.91 (0.75, 1.11)	0.3471	**1.28 (1.15, 1.43)**	**<.0001**
Hispanic ethnicity	1.13 (0.92, 1.39)	0.2510	1.10 (0.97, 1.24)	0.1558

*cell intentionally left blank

Questions for the Committee:

- Is there a gap in care that warrants a national performance measure?

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

1c. Composite – Quality Construct and Rationale

Maintenance measures – same emphasis on quality construct and rationale as for new measures.

1c. Composite Quality Construct and Rationale. The quality construct and rationale should be explicitly articulated and logical; a description of how the aggregation and weighting of the components is consistent with the quality construct and rationale also should be explicitly articulated and logical.

- The developer's rationale for the composite is that differentiating performance based on mortality alone fails to account for the fact that not all operative survivors received equal quality care. By combining the results of risk-adjusted mortality and the risk-adjusted occurrence of any of five major complications, this composite provides a more comprehensive quality assessment that should help participants identify potential areas for improvement. By aggregating the surgeries and rates, the

composite yields a more comprehensive view of participant performance, which may be more useful for accountability purposes.

- The developer notes that this measure is constructed using two domains:
    - Domain 1 is the absence of operative mortality (before hospital discharge or within 30 days of operation) for patients undergoing MVRR + CABG**.** This domain is calculated as a single measure.
    - Domain 2 is the absence of major morbidity, which is a "none or any" measure of the following complications: (1) prolonged ventilation; (2) deep sternal wound infection; (3) permanent stroke; (4) renal failure; and (5) reoperations for bleeding, prosthetic or native valve dysfunction, or other cardiac reasons, but not for other non-cardiac reasons.
    - The developer states that the domains are "rolled up" into a single number. They do not provide additional details on the weighting of the components in this section. In the Scientific Acceptability section (2d2.1), the developer states that the domains were rescaled by dividing their respective standard deviation across STS participants and then added together. After the rescaling, the relative weights were 0.74 for mortality and 0.26 for morbidity. The developer states that this weighting was consistent with their Expert Panel's clinical assessment of each domain's relative importance.

Questions for the Committee:

- Are the quality construct and a rationale for the composite explicitly stated and logical?
- Is the method for aggregation and weighting of the components explicitly stated and logical?

Preliminary rating for composite quality construct and rationale:
☐ **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.

Evidence continues to support.

no new evidence

No changes to evidence.

Evidence is acceptable.

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?

Some performance gap and disparities noted.

no concerns

Disparities shown based on type of insurance

Performance Gap – This measure has the largest performance gap of the 6 STS measures under review. Mean 86%, min = 77% and max of 92%, IQR of .33. The disparities data indicate that black patients are at higher risk of morbidity. But in 2015, only 10 participants (2.9%) had lower-than-expected performance

1c. Composite Performance Measure - Quality Construct (if applicable): Are the following stated and logical: overall quality construct, component performance measures, and their relationships; rationale and distinctive and additive value; and aggregation and weighting rules?

Quality construct and rationale seem appropriate.

expand information on the weighting rules as domains are rolled into a single number

Domains weighting according to expert panel assessment. Construct explained adequately.

The rationale for the composite is good. I am still unclear how much variation exists in each morbidity subcomponent

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability: Specifications and Testing

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

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

Reliability

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

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

Validity

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

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

Composite measures only:

2d. Empirical Analysis To Support Composite Construction. Empirical analysis should demonstrate that the component measures add value to the composite and that the aggregation and weighting rules are consistent with the quality construct.

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

Evaluators: NQF Staff

Scientific Acceptability Review

Questions for the Committee regarding reliability:

- Do you have any concerns that the measure can be consistently implemented (i.e., are measure specifications adequate)?
- The staff 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 staff raised concerns regarding the validity testing for the measure. What are your thoughts regarding the validity of the measure (e.g., exclusions, risk-adjustment approach, etc.)?

Questions for the Committee regarding composite construction:

- Do you have any concerns regarding the composite construction approach (e.g., do the component measures fit the quality construct and add value to the overall composite? Are the aggregation and weighting rules consistent with the quality construct and rationale while achieving the related objective of simplicity to the extent possible?)?
- The staff is satisfied with the composite construction. Does the Committee think there is a need to discuss and/or vote on the composite construction approach?

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

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

Preliminary rating for composite construction: ☐ **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?

No concerns with reliability specifications.

Are all data elements consistently pulled and reported in same manner (EHR capability vs abstract)

No issues.

Specifications are fine. Why didn't the SMP review this complex measure?

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

No concerns with reliability testing.

no

Testing adequate.

The method of calculating reliability using Bayesian estimate of true scores is more compelling in my opinion compared to the now common beta-binomial SNR. At the planned public reporting threshold of 25 index cases, the reliability is 0.50 (95% Prl 0.44 – 0.57). I consider this very modest reliability. Especially given the transformation of the measure scores into star ratings or 3-level categories, the more important analysis is the reliability of the below average classifications. I think this could be done with the same approach or a split sample method.

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

No data element testing.

no

Appears adequate for the described measure.

In known groups analysis, developers reported that compared to participants receiving 1 star, those with 3 stars had lower risk-adjusted mortality (3.0% vs. 11.2%) and lower risk-adjusted morbidity (20.9% vs. 52.3%) during July 2011 – June 2014. Test-retest reliability is not directly relevant to validity evidence. Another approach to validity testing might examine the relationship between the composite and structures/processes that are hypothesized to drive outcomes. The current logic model does not emphasize these relationships. These analyses could establish that measures scores are influenced by factors under the control of the participants and perhaps guide QI efforts.

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?

No issues.

no concerns

None

No problems

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

No major concerns.

missing data could pose a threat

none

no comment

2c. Composite Performance Measure - Composite Analysis (if applicable): Do analyses demonstrate the component measures fit the quality construct and add value? Do analyses demonstrate the aggregation and weighting rules fit the quality construct and rationale?

No issues to composite construction.

no concerns

Composite construction well defined

It is unclear how much each component adds to the composite. It would be helpful to expand the table that shows the average performance of the 2 main components across the deciles of composite scores, but unpacking the morbidities component into its sub-components.

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 data source for this measure is the STS registry. The STS database has more than 1,030 participants as of August 2020. Data are collected or generated and used by healthcare personnel during provision of care. Some institutions have full EHR capability; some may have partial or no availability. Some data elements are in defined fields in electronic sources, and some must be abstracted. However, all data from participating institutions are submitted in electronic format following standard data specifications.
- STS registry participants pay annual fees of $3,500 to $4,750 depending on whether the majority of surgeons in the group are STS members. There is an additional fee of $150 per member and $350 per non-member for surgeons listed on the database's Participation Agreement. Most participants also purchase data-entry software to submit data elements.

Questions for the Committee:

- Do you agree that the required data elements routinely generated and used during care delivery?
- Are the required data elements available in electronic form, e.g., EHR or other electronic sources?

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?

No issues or concerns with feasibility.

Due to small number 25 over 3 years)of submissions with measure, is this adequate to put measure into operational use?

Data elements readily available from STS registry and electronic sources.

The measure has been in use for years and appears 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

- The composite is publicly reported through the STS Public Reporting Program.

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

Feedback on the measure by those being measured or others

1. All Adult Cardiac Surgery Database participants receive quarterly feedback reports providing a detailed analysis of the participant's performance, including benchmarking. Dashboard-type reporting on STS.org has been provided for real-time, online data updates to STS surgeon members. Participants also have access to a guide to help interpret performance results.
2. The adult cardiac surgeons from across the U.S. who comprise the STS Adult Cardiac Surgery Task Force meet periodically to discuss the participant reports and to consider potential enhancements to the ACSD. This feedback was one of the drivers for the real-time dashboard-type reporting recently implemented.
3. The developer did not provide any examples of feedback being considered when changes are incorporated into the measure.

Additional Feedback:

Questions for the Committee:

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

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

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

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

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

Improvement results

The developer states that there has been a decrease in 1-star and 3-star ratings over time, which they state is consistent with their quality goal of reducing variation among participants.

Star ratings in percentages, 2017-2019

Stars	2019	2018	2017
*	2.55	2.08	2.74
**	88.0	89.97	91.78

Stars	2019	2018	2017
***	9.45	7.96	5.48

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

- None reported.

Potential harms

- Potential harms include gaming and risk aversion. The developer states that they control for these through a careful audit process and a robust risk-adjustment methodology.

Questions for the Committee:

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

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

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

4a1. Use - Accountability and Transparency: How is the measure being publicly reported? Are the performance results disclosed and available outside of the organizations or practices whose performance is measured? For maintenance measures - which accountability applications 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?

No concerns with use; publicly reported as part of accountability program. Feedback provided and received for the measure, no clear indication of incorporation in measure development.

no concerns

All Adult Cardiac surgery participants receive quarterly feedback. Task force meets regularly to review feedback and improve dashboard.

The measure is in use

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.

Shows improvement over time. No issues with usability.

no concerns

Decrease in 1 star rating minimal. Increase shown in 3 star ratings. 2 star ratings (average) remain similar. There is room for improvement in this measure.

Usability is probably good. Examining the relationship between the composite (and subcomponents) and structures/processes that are hypothesized might inform participants how they can drive improvements

Criterion 5: Related and Competing Measures

Related or competing measures
NQF #0696 STS CABG Composite
NQF #2561 Aortic Valve Replacement Composite Score
NQF #2563 Aortic Valve Replacement + CABG Composite Score
NQF #3031 Mitral Valve Repair/Replacement Composite Score

Harmonization
The identified measures are all developed by STS and the developer indicates that they are harmonized.

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?

Few related measures, no concerns with harmonization.

no concerns

All competing measures developed by STS and are harmonized.

no comments

Public and Member Comments

Comments and Member Support/Non-Support Submitted as of: 01/26/2021

Comment by: Society of Thoracic Surgeons

STS Response to Preliminary Analyses for Measures 3030, 3031, 3032: "Insufficient" ratings for Validity

For each of these composite measures, the Preliminary Analysis states that "Demonstrating a relationship between performance on the overall composite and the composite domains may not be a valid assessment of composite score validity." As in past endorsement and endorsement maintenance reviews for our composite measures, we believe it to be a reasonable approach to use our morbidity and mortality component scores as the "gold standard" against which to demonstrate construct or criterion validity of the composite scores across our three performance categories: "higher-than-expected," "lower-than-expected," and "as-expected" (as defined in 2b1.2 in our composite testing forms). If participants/surgeons with "higher-than-expected" composite ratings have consistently lower risk-adjusted mortality and lower risk-adjusted morbidity compared to participants/surgeons with "lower-than-expected" ratings, we believe the validity of the composite score is demonstrated. The STS has the most sophisticated outcomes data and methodology available for heart surgery, in a database with over 95% penetration across cardiac surgery practices in the U.S.; we therefore have no other "gold standard" against which to compare our results.

NQF staff have suggested the use of an external standard – e.g., a measure for a different cardiothoracic surgery procedure – for testing the validity of our composite measures. However, published studies have shown that excellent performance on one surgical procedure does not necessarily correlate with excellent performance on another procedure. We therefore maintain that the approach described above is appropriate for demonstrating the validity of our composite measures.

STS Updates to Measure Testing Document Section 1.8 for Measures NQF#s 3030, 3031, 3032 - PART 1

1.8 What were the social risk factors that were available and analyzed?

The STS position on inclusion of social risk factors (e.g., SES/SDS/race) as risk model variables is best summarized in this excerpt from our 2018 risk model publication [1]. We describe in detail the controversies about such variables, and how we have attempted to reconcile them:

"Whether outcomes measures, and the public reporting and reimbursement programs based on them, should consider socioeconomic status (SES) or sociodemographic factors (e.g., race, ethnicity, education, income, payer [e.g., Medicare-Medicaid dual eligible status]) is a topic of intense health policy debate. Some argue that in the absence of adjustment for these variables, the outcomes of hospitals that care for a disproportionate percentage of low SES patients will be unfairly disadvantaged, perhaps leading to financial or reputational penalties. Opponents argue that inclusion of SES factors in risk models may "adjust away" disparities in quality of care, and they advocate the use of stratified analyses instead. Also, readily available SES factors have often not demonstrated significant impact on outcomes, perhaps because they are not sufficiently granular or relevant. Finally, even SES proponents agree that these factors make more sense conceptually for some outcomes (e.g., readmission) than for others (hospital mortality, complications). Notably, as part of a National Quality Forum pilot project, the STS specifically studied dual eligible status in the STS readmission measure and found minimal impact. In developing the new STS risk models, we avoided these more philosophical and health policy arguments regarding SES adjustment and based our modeling decisions on empiric findings and consideration of the model's primary intended purpose—optimal case mix adjustment. Conceptually, our goal was to adjust for all preoperative factors that are independently and significantly associated with outcomes and that vary across STS participants. For example, race will continue to be in our risk models as it has been previously, but not conceptually as a SES indicator [Note: nor as a surrogate for such factors]. Race has an empiric association with outcomes and has the potential to confound the interpretation of a hospital's outcomes, although we do not know the underlying mechanism (e.g., genetic factors, differential effectiveness of certain medications, rates of certain associated diseases such as diabetes and hypertension, and potentially SES for some outcomes such as readmission)."

STS is aware of the recent NEJM paper by Vyas and colleagues [2] and has directly communicated with the lead author to explain why race is included in STS models, and to correct several misinterpretations and misrepresentations in this article. Dr. Vyas acknowledged that they included extended quotes from our risk model paper precisely because we were one of the few risk model developers that thoroughly described our rationale for race inclusion, as noted in the excerpt above.

Documents produced by NQF [3, 4], the National Academy of Medicine [5-8], the Office of the Assistant Secretary for Planning and Evaluation (Social Risk Factors and Performance Under Medicare's Value-Based Purchasing Programs) [9], and as part of the 21st Century Cures Act legislation [10] are particularly instructive. They summarize the arguments for and against inclusion of SDS/SES/racial adjustment in risk models; context-specific considerations for when they might be appropriate or inappropriate; strategies to avoid the potential adverse unintended consequences of such adjustment; concomitant monitoring for social and racial inequities through stratification; and special approaches for providers who care for high proportions of disadvantaged populations (e.g., payment adjustments, additional resources).

Adjustment for SDS/SES/racial factors has generally been regarded as acceptable (e.g., in NQF white papers) when there is both an empirical association AND a plausible conceptual association of the risk variable with an outcome. For example, an SES/SDS/racial risk factor might be appropriate as a risk variable for readmission or mortality risk models, but not for CAUTI (catheter-associated urinary tract infections), CLABSI (central line-associated bloodstream infection), or process measures.

For many outcomes, SES/SDS/racial adjustment is warranted to optimize risk model accuracy. For example, recent STS and Duke Clinical Research Institute analyses show that if race variables are excluded from some STS models, the resulting outcomes estimates are markedly different than the actual observed outcomes, and the O/E ratios are significantly different than unity, especially when the models are applied to racial minority subpopulations—in other words, the models are less well calibrated, an essential feature of any risk model.

This miscalibration persisted even when an SES/SDS indicator (specifically, dual eligible status) was simultaneously included in the models (i.e., thus addressing the hypothesis that the putative association of race and various outcomes is actually mediated by SES/SDS). Use of risk estimates from such models for patient counseling and shared decision-making would be misleading to patients and would inaccurately portray (and unfairly disadvantage) the risk-adjusted performance of providers, especially those caring for minority populations. Importantly, STS and its analytic center re-estimate risk factor coefficients several times annually, so that any changes in the association of race with outcomes will be implemented in the newest estimates. Further, STS is geocoding it adult cardiac surgery records and will use this information to derive an Area Deprivation Index for all patients with a valid address, thus providing us with the ability to further study the impact of race and SES/SDS using what is arguably the most sensitive and comprehensive SES/SDS indicator. Finally, STS is aware of the recommendation in the ASPE report of October 2020 that functional status indicators be included in risk models as it may account for some of the impact on outcomes associated that is currently attributed to race. Although STS has a well-documented frailty indicator (5 meter walk test), it has not been collected with sufficient consistency by our participants to allow its inclusion in our models. Accordingly, STS has established a new working group on Frailty/functional indicators whose goal is to develop a new indicator that can be captured for virtually all patients using a combination of history, lab data, functional status, etc. Once developed, it will be added to STS models.

Although SDS/SES/racial risk adjustment may be indicated to assure optimal risk model estimates based on current data, it is widely believed that such adjustment could potentially obscure disparities in care. To avoid this potential unintended consequence, most of the national guidance documents cited above recommend that any risk model results that are adjusted for SES/SDS/racial factors also present concomitant results in which outcomes are stratified by the same variables. This is a much more direct and explicit approach to monitor disparities and inequities and has been followed by STS in its risk modeling and performance measures. Please refer to the race-specific disparities data provided for each of the domains (mortality and morbidity) of measure 3030 under question 1b.4 (Importance tab) of the submission form (to be completed by the November submission deadline), which we believe will suffice to comply with this recommendation.

STS Updates to Measure Testing Document Section 1.8 for Measures NQF #s 3030, 3031, 3032 - PART 2

1.8 What were the social risk factors that were available and analyzed?

1. Shahian DM, Jacobs JP, Badhwar V, Kurlansky PA, Furnary AP, Cleveland JC, Jr., et al. The Society of Thoracic Surgeons 2018 Adult Cardiac Surgery Risk Models: Part 1-Background, Design Considerations, and Model Development. Ann Thorac Surg. 2018;105(5):1411-8.
2. Vyas DA, Eisenstein LG, Jones DS. Hidden in Plain Sight — Reconsidering the Use of Race Correction in Clinical Algorithms. New England Journal of Medicine. 2020.
3. National Quality Forum. Risk adjustment for Socioeconomic Status or other Sociodemographic Factors, accessed at http://www.qualityforum.org/Publications/2014/08/Risk_Adjustment_for_Socioeconomic_Status_or_Other_Sociodemographic_Factors.aspx on June 24, 2020. 2014.
4. The National Quality Forum. Evaluation of the NQF Trial Period for Risk Adjustment for Social Risk Factors. January 15, 2017. Available from: https://www.qualityforum.org/Publications/2017/07/Social_Risk_Trial_Final_Report.aspx.
5. National Academies of Sciences, Engineering, and Medicine. Accounting for social risk factors in Medicare payment. Washington, DC: The National Academies Press; 2017.
6. National Academies of Sciences, Engineering, and Medicine. Accounting for social risk factors in Medicare payment: Data. Washington, DC; 2016.
7. National Academies of Sciences, Engineering, Medicine. Accounting for Social Risk Factors in Medicare Payment: Criteria, Factors, and Methods. Washington, DC: The National Academies Press; 2016.

8. National Academies of Sciences, Engineering, Medicine,. Accounting for Social Risk Factors in Medicare Payment: Identifying Social Risk Factors. Washington, DC: The National Academies Press; 2016. 110 p.
9. Office of the Assistant Secretary for Planning and Evaluation USDoHaHS. Report to Congress: Social Risk Factors and Performance Under Medicare's Value-Based Purchasing Programs. A Report Required by the Improving Medicare Post-Acute Care Transformation (IMPACT) Act of 2014. Washington, DC; 2016.
10. 114th Congress of the United States. 21st Century Cures Act (Public Law 114–255). Washington, DC; 2016.

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

Scientific Acceptability Evaluation

Scientific Acceptability: Preliminary Analysis Form

Measure Number: 3032

Measure Title: STS Mitral Valve Repair/Replacement + Coronary Artery Bypass Graft (CABG)

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.**
   - No concerns that STS can consistently implement and calculate the measure.

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

   - The developer conducted one set of testing for clinician group and facility. For the adult cardiac database, 92% of the participants are surgical groups with a one-to-one relationship to an individual facility.
   - The developer conducted composite-score-level signal-to-noise analysis. They utilized a Bayesian approach to generate possible values for each participant's score and then estimated the true values by conducting Markov Chain Monte Carlo simulations. The data used in the simulation are from a three-year period of July 2011 – June 2014, which is rather dated. The developer included results for a range of case counts and indicate that they intend to use a 25-case threshold for public reporting.
7. **Assess the results of reliability testing**

   Submission document: Testing attachment, section 2a2.3

   - The results range from a reliability of 0.42 (95% Prl 0.0.35 – 0.0.48) to 0.62 (95% Prl 0.52 – 0.70) for 50 cases. At the planned public reporting threshold of 25 index cases, the reliability is 0.0.50 (95% Prl 0.44 – 0.57). This demonstrates moderate 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)
10. **OVERALL RATING OF RELIABILITY** (taking into account precision of specifications and <u>all</u> testing results):

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

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

    ☐ **Low** (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.**
    - Precise specifications (Box 1) → Empiric reliability testing (Box 2) → Testing at measure score level (Box 4) → Method described and appropriate (Box 5) → Level of confidence (Box 6) →Moderate

VALIDITY: ASSESSMENT OF THREATS TO VALIDITY

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

    Submission document: Testing attachment, section 2b2.
    - The developers indicate that this measure has no exclusions.

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

    Submission document: Testing attachment, section 2b4.
    - No concerns.
    - Among surgeons with at least 25 cases over 3 years (July 2012 – June 2015), around 91% of participants received 2 stars, and the remaining participants received either 1 or 3 stars.
        - 314 (91.0%) performed as expected
        - 10 (2.9%) had lower-than-expected performance
        - 21 (6.1%) had higher-than-expected performance

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.
    - No concerns. There is only one data source/method for this measure.

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

    Submission document: Testing attachment, section 2b6.
    - 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

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

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

    16d. **Risk adjustment summary:**

    16d.1 All of the risk-adjustment variables present at the start of care? ☒ Yes ☐ No

    16d.2 If factors not present at the start of care, do you agree with the rationale provided for inclusion? ☐ Yes ☐ No

    16d.3 Is the risk adjustment approach appropriately developed and assessed? ☒ Yes ☐ No

    16d.4 Do analyses indicate acceptable results (e.g., acceptable discrimination and calibration)

☒ Yes ☐ No

16d.5. Appropriate risk-adjustment strategy included in the measure? ☒ Yes ☐ No

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

- To adjust for case mix in this measure, the developer modified and re-estimated the previously published 2008 STS valve+CABG model. The need for modification was due to broader inclusion criteria for this measure and to account for the major morbidity component.
- The bootstrap-adjusted estimated c-statistic was 0.708 for the morbidity model and 0.738 for the mortality model. The developer interprets this to demonstrate well calibrated risk models with good discrimination power.

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

- Measure score validity was examined using known-groups validity. Participants were divided into three groups as follows:
  - Participants were labeled as having higher-than-expected performance if the 95% credible interval surrounding a participant's composite score fell entirely above the overall STS average composite score.
  - Participants were labeled as having lower-than-expected performance if the 95% credible interval surrounding a participant's composite score fell entirely below the overall STS average composite score.
  - Participants were labeled as higher-than-expected performance (3 stars), lower-than-expected performance (1 star), and indistinguishable from the average or as-expected performance (2 stars).
- Mortality (domain 1) and morbidity (domain 2) scores were then compared for each group of participants.
- Demonstrating a relationship between performance on the overall composite and the composite domains may not be a valid assessment of composite score validity.
- Measure score validity was also examined using predictive validity/stability of measure score results over time. Stability could be considered a test of reliability vs a test of validity of a measure. This methodology has been accepted to demonstrate validity in previous submissions

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

Submission document: Testing attachment, section 2b2.3

- The developers reported that compared to participants receiving 1 star, those with 3 stars had lower risk-adjusted mortality (3.0% vs. 11.2%) and lower risk-adjusted morbidity (20.9% vs. 52.3%) during

July 2011 – June 2014. Thus, differences in performance were clinically meaningful as well as statistically significant. STS participants deemed better by the composite scores have (on average) higher performance during the same time window on each individual domain of the composite measure.

- Demonstrating a relationship between performance on the overall composite and the composite domains may not be a valid assessment of composite score validity.
- For the data periods July 2011 – June 2014 and July 2012 – June 2015 the Pearson correlation between composite scores was 0.79.

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

    Submission document: Testing attachment, section 2b1.

    ☐ **Yes**

    ☒ **No**

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

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

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

    ☐ **Yes**

    ☐ **No**

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

25. **OVERALL RATING OF VALIDITY 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.**

    The information and testing provided is not sufficient to determine the validity of the composite measure.

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**

- Pearson correlations were calculated to verify that each of the two domains of the measure contribute statistical information but do not dominate the composite. Data from July 2011 – June 2014 were used for the calculation. Results were 0.60 for mortality domain versus overall composite measure and 0.91 for morbidity domain score versus overall score. The developers interpret this to mean that risk-adjusted morbidity explains more of the variation in the overall composite score but does not dominate the score.
- The developer states that the domains were rescaled by dividing their respective standard deviation across STS participants and then the domains were added together. After the rescaling, the relative weights were 0.74 mortality and 0.26 morbidity.
- Weighting was assessed by an expert panel. It was consistent with the panel's clinical assessment of each domain's relative importance. The developer states that a one percentage point change in a participant's risk-adjusted mortality rate has the same impact on the overall score as a 2.8 percentage point change in the site's risk-adjusted morbidity rate.

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.**

Developer Submission

NQF #: 3032

Corresponding Measures:

De.2. Measure Title: STS Mitral Valve Repair/Replacement (MVRR) + Coronary Artery Bypass Graft (CABG) Composite Score

Co.1.1. Measure Steward: The Society of Thoracic Surgeons

De.3. Brief Description of Measure: The STS Mitral Valve Repair/Replacement (MVRR) + Coronary Artery Bypass Graft (CABG) Composite Score measures surgical performance for MVRR + CABG with or without concomitant Atrial Septal Defect (ASD) and Patient Foramen Ovale (PFO) closures, tricuspid valve repair (TVr), or surgical ablation for atrial fibrillation (AF). To assess overall quality, the STS MVRR +CABG Composite Score comprises two domains consisting of six measures:

Domain 1 – Absence of Operative Mortality

Proportion of patients (risk-adjusted) who do not experience operative mortality. Operative mortality is defined as death before hospital discharge or within 30 days of the operation.

Domain 2 – Absence of Major Morbidity

Proportion of patients (risk-adjusted) who do not experience any major morbidity. Major morbidity is defined as the occurrence of any one or more of the following major complications:

1. Prolonged ventilation,
2. Deep sternal wound infection,
3. Permanent stroke,
4. Renal failure, and
5. Reoperations for bleeding, coronary graft occlusion, prosthetic or native valve dysfunction, and other cardiac reasons, but not for other non-cardiac reasons.

Outcome data are collected on all patients and from all participants. For optimal measure reliability, participants meeting a volume threshold of at least 25 cases over 3 years receive a score for each of the two domains, plus an overall composite score. The overall composite score is created by “rolling up” the domain scores into a single number. In addition to receiving a numeric score, participants are assigned to rating categories designated by the following:

1 star – lower-than-expected performance

2 stars – as-expected performance

3 stars – higher-than-expected performance

1b.1. Developer Rationale: N/A

S.4. Numerator Statement: Due to the complex methodology used to construct the composite measure, it is impractical to separately discuss the numerator and denominator. The following discussion describes how each domain score is calculated and how these are combined into an overall composite score.

The STS Mitral Valve Repair/Replacement (MVRR) Composite Score comprises two domains consisting of six measures:

Domain 1 – Absence of Operative Mortality

Proportion of patients (risk-adjusted) who do not experience operative mortality. Operative mortality is defined as death before hospital discharge or within 30 days of the operation.

Domain 2 – Absence of Major Morbidity

Proportion of patients (risk-adjusted) who do not experience any major morbidity. Major morbidity is defined as the occurrence of any one or more of the following major complications:

1. Prolonged ventilation,
2. Deep sternal wound infection,
3. Permanent stroke,
4. Renal failure, and
5. Reoperations for bleeding, coronary graft occlusion, prosthetic or native valve dysfunction, and other cardiac reasons, but not for other non-cardiac reasons.

Participants receive a score for each of the two domains, plus an overall composite score. The overall composite score was created by "rolling up" the domain scores into a single number. In addition to receiving a numeric score, participants are assigned to rating categories designated by the following:

1 star – lower-than-expected performance

2 stars – as-expected performance

3 stars – higher-than-expected performance

Patient Population: The analysis population consists of patients aged 18 years or older who MVRR + CABG with or without concomitant Atrial Septal Defect (ASD) and Patient Foramen Ovale (PFO) closures, tricuspid valve repair (TVr), or surgical ablation for atrial fibrillation (AF).

Time Window: 3 years

Data Completeness Requirement: Participants are excluded from the analysis if they have fewer than 25 MVRR + CABG procedures in the patient population.

Estimation of Composite Scores and Star Ratings:

To be consistent with the conventions of previous composite measures, risk-adjusted event rates were first converted into risk-adjusted absence-of-event rates. To calculate the composite, participant-specific absence of mortality rates and absence of morbidity rates were weighted inversely by their respective standard deviations across participants. This procedure was equivalent to first rescaling the absence of mortality rates and absence of morbidity rates by their respective standard deviations across participants, and then assigning equal weighting to the rescaled rates. Finally, in order to draw statistical inferences about participant performance, a Bayesian credible interval surrounding each participant's composite score was calculated. Unlike frequentist confidence intervals, Bayesian credible intervals have an intuitively direct interpretation as an interval containing the true value of the composite score with a specified probability (e.g., 95%). To determine star ratings for each participant, the credible interval of its composite score was compared with the STS average. Participants whose intervals were entirely above the STS average were classified as 3-star (higher than expected performance), and participants whose intervals were entirely below the STS average were classified as 1-star (lower than expected performance). Credible intervals based on different probability levels (90%, 95%, 98%) were explored, and the resulting percentages of 1, 2, and 3-star programs were calculated.

S.6. Denominator Statement: See response in S.4. Numerator Statement for complete description of measure specifications.

Patient Population: The analysis population consists of patients aged 18 years or older who MVRR + CABG with or without concomitant Atrial Septal Defect (ASD) and Patient Foramen Ovale (PFO) closures, tricuspid valve repair (TVr), or surgical ablation for atrial fibrillation (AF).See response in S.4. Numerator Statement for complete description of measure specifications.

Patient Population: The analysis population consists of patients aged 18 years or older who MVRR + CABG with or without concomitant Atrial Septal Defect (ASD) and Patient Foramen Ovale (PFO) closures, tricuspid valve repair (TVr), or surgical ablation for atrial fibrillation (AF).

S.8. Denominator Exclusions: Data Completeness Requirement: Participants are excluded from the analysis if they have fewer than 25 MVRR + CABG procedures in the patient population.

De.1. Measure Type: Composite

S.17. Data Source: Registry Data

S.20. Level of Analysis: Clinician : Group/Practice, Facility

IF Endorsement Maintenance – Original Endorsement Date: Jan 25, 2017 **Most Recent Endorsement Date:** Jan 25, 2017

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

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

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

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

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

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

7.1-Evidence_Form-3032-STS_MVRR-CABG_Comp_Score-Fall2020.docx

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

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

No

1a. Evidence (subcriterion 1a)

NATIONAL QUALITY FORUM—Evidence (subcriterion 1a)

Measure Number (*if previously endorsed*)**:** 3032

Measure Title: STS Mitral Valve Repair/Replacement (MVRR) + CABG Composite Score

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

Date of Submission: **11/16/2020**

Instructions

- Complete 1a.1 and 1a.2 for all measures. If instrument-based measure, complete 1a.3.
- Complete EITHER 1a.2, 1a.3 or 1a.4 as applicable for the type of measure and evidence.
- For composite performance measures:
  - A separate evidence form is required for each component measure unless several components were studied together.

- If a component measure is submitted as an individual performance measure, attach the evidence form to the individual measure submission.

- All information needed to demonstrate meeting the evidence subcriterion (1a) must be in this form. An appendix of supplemental materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.

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

1a. Evidence to Support the Measure Focus

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

- Outcome: [3] Empirical data demonstrate a relationship between the outcome and at least one healthcare structure, process, intervention, or service. If not available, wide variation in performance can be used as evidence, assuming the data are from a robust number of providers and results are not subject to systematic bias.
- Intermediate Clinical Outcome: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured intermediate clinical outcome leads to a desired health outcome.
- Process: [5] a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured process leads to a desired health outcome.
- **Structure**: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured structure leads to a desired health outcome.
- **Efficiency**: [6] evidence not required for the resource use component.
- For measures derived from **patient reports**, evidence should demonstrate that the target population values the measured outcome, process, or structure and finds it meaningful.
- **Process measures incorporating Appropriate Use Criteria:** See NQF's guidance for evidence for measures, in general; guidance for measures specifically based on clinical practice guidelines apply as well.

Notes

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

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

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

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

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

Outcome

☒ Outcome: 1. Operative Mortality; 2. Postoperative Major Morbidity

☐ 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.

Operative Mortality

The incidence of mitral valve incompetence resulting from or coexistent with coronary artery disease is increasing. This results from a progressively older population of patients, with recommendations for earlier surgical intervention, among other factors. Patients undergoing combined coronary bypass and mitral valve surgery have an increasing number and severity of co-morbid risk factors. As a result, these patients have among the highest mortality rate of all standard cardiac surgical procedures. Mortality is likely the single most important negative outcome that can be associated with a surgical procedure. Critical evaluation of operative mortality allows one to evaluate the risk associated with a given procedure for various patient characteristics, and more importantly, aggressively search for ways to minimize that risk.

Major Morbidity

- **Prolonged Ventilation.** Prolonged ventilation has been shown to substantially increase length of stay and cost of care, and is associated with higher rates of respiratory failure, stroke, renal failure, and death. Modalities to decrease the rate of prolonged intubation include physician-supervised extubation protocols implemented by nurses and respiratory therapists, improved preoperative preparation of patients, reduction of postoperative bleeding, and intra-operative protocolized anesthesia care. Current implementation is highly variable and great opportunities exist to increase the implementation of evidence-based care. Cardiac surgery programs with high implementation have lower than average rates of prolonged ventilation, resulting in significantly lower rates of adverse events.
- **Deep sternal wound infection.** A devastating complication of cardiac surgery is deep sternal wound infection. Patients with deep sternal wound infection require multiple surgeries to clear the infection, and have longer hospital stays, greatly increased costs and increased early and late mortality. Care processes that influence the incidence of sternal infection span the first 3 major phases of surgical care. In the preoperative phase, routine patient decontamination and identification of active infections are key elements. In the intraoperative phase, impeccable surgical aseptic technique and timing of antibiotic administration are key processes. In the postoperative phase, aseptic wound care and detection of early wound inflammation are important in preventing delayed contamination and subsequent infection.
- **Permanent stroke.** Postoperative stroke/CVA results in significant short- and long-term potentially devastating effects for patients and their families. It is associated with significant increases in death, respiratory failure, renal failure, length of stay, and cost of care. Modalities to decrease the rate of postoperative stroke include measures to maintain blood pressure and cerebral perfusion, glycemic

control, avoidance of atrial fibrillation, anticoagulation protocols, careful intraoperative imaging, etc. Many opportunities exist to decrease stroke rates by increasing implementation of evidence-based strategies.

- **Renal failure.** Postoperative renal failure is an occasional but serious complication in the cardiac surgical population and is a major determinant of short- and long-term survival. Identification of clinical precursors of postoperative renal insufficiency and improvement in perioperative management of this high-risk group will improve the long-term survival of these patients. By implementing known recommendations (delay heart surgery, when possible, after cardiac catheterization, maintain mean CPB perfusion pressure at 80% of preop BP, etc.), postoperative kidney injury should be significantly reduced.
- **Reoperation.** Surgical re-exploration for bleeding remains a known complication following cardiac surgery. The literature documents that bleeding following coronary artery bypass surgery portends a longer ICU stay and therefore greater resource consumption. It remains controversial whether long-term outcomes are worse for patients whose reoperation is re-exploration for bleeding. However, Hein *et al* document that patients with an ICU stay > 3 days (with bleeding as a multivariate risk factor for this outcome) have a long-term survival which is inferior to patients with an ICU stay < 3 days. Patients who undergo reoperation for other cardiac reasons likewise have a longer ICU length of stay and higher mortality.

References – Operative Mortality

- Birkmeyer NJO, Marrin CA, et al. Decreasing mortality for aortic and mitral valve surgery in Northern New England. Northern New England Cardiovascular Disease Study Group. Ann Thorac Surg 2000;70(2):432-437.
- Edwards FH, Peterson ED, et al. Prediction of operative mortality following valve replacement surgery. J Am Coll Cardiol 2001;37:885-92.
- Goodney PP, O'Connor GT, et al. Do hospitals with low mortality rates in coronary artery bypass also perform well in valve replacement? Ann Thorac Surg 2003;76:1131-7.
- Mehta RH, Eagle KA, et al. Influence of age on outcomes in patients undergoing mitral valve replacement. Ann Thorac Surg 2002;74:1459-67.
- Shahian DM, O'Brien SM, Filardo G, Ferraris VA, Haan CK, Rich JB, Normand SL, DeLong ER, Shewan CM, Dokholyan RS, Peterson ED, Edwards FH, Anderson RP. The Society of Thoracic Surgeons 2008 cardiac surgery risk models: part 3--valve plus coronary artery bypass grafting surgery. Ann Thorac Surg 2009;88(1 Suppl):S43-62.
- Miyata H, Motomura N, Tsukihara H, Takamoto S, Japan Cardiovascular Surgery Database. Risk models including high-risk cardiovascular procedures: clinical predictors of mortality and morbidity. Eur J Cardiothorac Surg 2011;39:667-74.
- Vassileva CM, Boley T, Markwell S, Hazelrigg S. Meta-analysis of short-term and long-term survival following repair versus replacement for ischemic mitral regurgitation. Eur J Cardiothorac Surg 2011; 39:295-303.
- Daneshmand MA, Milano CA, Rankin JS, Honeycutt EF, Shaw LK, Davis RD, Wolfe WG, Glower DD, Smith PK. Influence of patient age on procedural selection in mitral valve surgery. Ann Thorac Surg 2010;90:1479-85.
- Acker MA, Parides MK, Perrault LP, Moskowitz AJ, Gelijns AC, Voisine P, Smith PK, Jung JW, Blackstone EH, Puskas JD, Argenziano M, Gammie JS, et al. Mitral-valve repair versus replacement for severe ischemic mitral regurgitation. N Engl J Med 2014;370:23-32.
- Badhwar V, Vemulapalli S, Mack MA, et al. Volume-outcome association of mitral valve surgery in the United States. JAMA Cardiol 2020;5:1092-101.

References – Major Morbidity

- Shroyer AL, Coombs LP, Peterson ED, et al. The Society of Thoracic Surgeons: 30-day operative mortality and morbidity risk models. Ann Thorac Surg 2003;75:1856-65.
- Hein OV, Birnbaum J, Wernecke K, England M, Knoertz W, Spies C. Prolonged Intensive Care Unit Stay in Cardiac Surgery: Risk Factors and Long-term Survival. Ann Thor Surg 2006;81:880-85.
- Stamou SC, Camp SL, Stiegel RM, et al. Quality improvement program decreases mortality after cardiac surgery. J Thorac Cardiovasc Surg 2008;136:494-99.
- Braxton JH, Marrin CA, McGrath PD, et al. 10-Year follow-up of patients with and without mediastinitis. Semin Thorac Cardiovasc Surg. 2004;16:70–76.
- Graf K, Ott E, Vonberg RP, et al. Economic aspects of deep sternal wound infections. Eur J Cardiothorac Surg 2010;37:893-96.
- Edwards FH, Engelman RM, Houck P et al. The Society of Thoracic Surgeons practice guideline series: antibiotic prophylaxis in cardiac surgery, part I: duration. Ann Thorac Surg 2006;81: 397-404,
- Wilson APR, Gibbons C, Reeves BC, et al. Surgical wound infection as a performance indicator: agreement of common definitions of wound infection in 4773 patients. BMJ 2004;329:720-4.
- Filsoufi F, Castillo JG, Rahmanian PB, et al. Epidemiology of deep sternal wound infection in cardiac surgery. J Cardiothorac Vasc Anesth 2009;23:488-94.
- Koch CG, Nowicki ER, Rajeswaran J, et al. When the timing is right: antibiotic timing and infection after cardiac surgery. J Thorac Cardiovasc Surg 2012;144:931-7.
- Paul M, Raz, A, Leibovici L, et al. Sternal wound infection after coronary artery bypass graft surgery: validation of existing risk scores. J Thorac Cardiovasc Surg 2007;133:397-403.
- Lazar HL, Ketchedjian A, Haime M, et al. Topical Vancomycin in combination with perioperative antibiotics and tight glycemic control helps to eliminate sternal wound infections. J Thorac Cardiovasc Surg 2014;148:1035-40.
- Miyahara K, Matsuura A, Takemura H, et al. Implementation of bundled interventions greatly decreases deep sternal wound infection following cardiovascular surgery. J Thorac Cardiovasc Surg 2014;148:2381-8.
- Matros E, Aranki, SF, Bayer LR, et al. Reduction in incidence of deep sternal wound infections: random or real? J Thorac Cardiovasc Surg 2010;139:680-5.
- Bardell T, Legare JF, Buth KJ, et al. ICU readmission after cardiac surgery. Eur J Cardiothorac Surg. 2003;23:354-9.
- Meade MO, Guyatt G, Butler R, et al. Trials comparing early vs late extubation following cardiovascular surgery. Chest 2001;120(6 Suppl):445S-53S.
- Naughton C, Reilly N, Powroznyk A, et al. Factors determining the duration of tracheal intubation in cardiac surgery: a single-centre sequential patient audit. Eur J Anaesthesiol. 2003;20:225-233.
- Welke KF, Ferguson TB, Coombs LP, et al. Validity of the Society of Thoracic Surgeons National Adult Cardiac Surgery Database. Ann Thorac Surg 2004;77:1137-39.
- Knapik P, Ciesla D, Borowik D, Czempik P, Knapik T. Prolonged ventilation post cardiac surgery – tips and pitfalls of the prediction game. J Cardiothorac Surg 2011;6:158.
- Alsabbagh MM, Asmar A, Ejaz NI, Aiyer RK, Kambhampati G, Ejaz AA. Update on clinical trials for the prevention of acute kidney injury in patients undergoing cardiac surgery. Am J Surg 2013;206:86-95.

- Arora P, Kolli, H, Nainani N, Nader N, Lohr J. Preventable risk factors for acute kidney injury in patients undergoing cardiac surgery. J Cardiothorac Vasc Anesth 2012;26:687-97.
- Boldt J, Brenner T, Lehmann A, Suttner SW, Kumle B, Isgro F. Is kidney function altered by the duration of cardiopulmonary bypass? Ann Thorac Surg 2003;75:906-12.
- Chertow GM, Levy EM, Hammermeister KE, et al. Independent association between acute renal failure and mortality following cardiac surgery. Am J Med 1998;104:343-348.
- Conlon PJ, Stafford-Smith M, White WD, Newman MF, King S, Winn MP, Landolfo K. Acute renal failure following cardiac surgery. Nephrol Dial Transplant 1999;14:1158-62.
- Kramer RS, Quinn RD, Groom RC, Braxton JH, Malenka DJ, Kellett MA, Brown JR for the Northern New England Cardiovascular Disease Study Group. Same admission cardiac catheterization and cardiac surgery: is there an increased incidence of acute kidney injury? Ann Thorac Surg 2010;90:1418-1424.
- Mangano CM, Diamondstone LS, Ramsay JG, et al. Renal dysfunction after myocardial revascularization: risk factors, adverse outcomes, and hospital resource utilization: the Multicenter Study of Perioperative Ischemia Research Group. Ann Intern Med 1998;128:194-203.
- Ranucci M, Ballotta A, Agnelli B, Frigiola A, Mencanti L, Castelvecchio S, for the Surgical and Clinical Outcome Research (SCORE) Group. Acute kidney injury in patients undergoing cardiac surgery and coronary angiography on the same day. Ann Thorac Surg 2013;95:513-9.
- Rosner MH, Okusa MD. Acute kidney injury associated with cardiac surgery. Clin J Am Soc Nephrol 2006;1:19-32.
- Shahian DM, Edwards FH, Ferraris VA, Haan CK, Rich JB, Normand SLT, DeLong ER, O'Brien SM, Shewan CM, Dokholyan RS, Peterson ED. Quality measurement in adult cardiac surgery: Part 1-conceptual framework and measure selection. Ann Thorac Surg 2007;83:S3-S12.
- Udesh R, Mehta A, Gleason TG, Wechsler L, Thirumala PD. Perioperative strokes and early outcomes in mitral valve surgery: a nationwide analysis. J Cardiothorac Vasc Anesth 2017;31:529-36.
- Tang AT, Alexiou C, Hsu J, Sheppard SV, Haw MP, Ohri SK. Leukodepletion reduces renal injury in coronary revascularization: a prospective randomized study. Ann Thorac Surg 2002;74:372-7.
- Amory DW, Grigore A, Amory JK, et al. Neuroprotection is associated with beta-adrenergic receptor antagonists during cardiac surgery: evidence from 1,575 patients. J Cardiothorac Vasc Anesth 2002;16:270-7.
- Arsenault KA, Yusus AM, Crystal E, Healey JS, Morillo CA, Nair GM, et al. Interventions for preventing postoperative atrial fibrillation in patients undergoing heart surgery. Cochrane Database Syst Rev 2013; 1:CD003611.
- Bucerius J, Gummert JF, Borger MA, et al. Predictors of delirium after cardiac surgery delirium: effect of beating-heart (off-pump) surgery. J Thorac Cardiovasc Surg 2004;127:57-64.
- Inoue K, Luth JU, Pottkamper D, et al. Incidence and risk factors of perioperative cerebral complications: heart transplantation compared to coronary artery bypass grafting and valve surgery. J Cardiovasc Surg 1998;39:201-8.
- Rosenberger P, Shernan SK, Loffler M, Shekar PS, Fox JA, Tuli JK, Nowak M and Eltzschig HK. The influence of epiaortic ultrasonography in intraoperative surgical management in 6051 cardiac surgical patients. Ann Thorac Surg 2008;85:548-53.
- Iribarne A, Burgener JD, Hong K, Raman J, Akhter S, Easterwood R, Jeevanandam V, Russo MJ. Quantifying the incremental cost of complications associated with mitral valve surgery in the United States. J Thorac Cardiovasc Surg 2012;143:864-72.

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

N/A

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

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

Please see response in 1a.2 (Logic Model) above.

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	*

Systematic Review	Evidence
not a guideline, summarize the conclusions from the SR.	
Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	*
Provide all other grades and definitions from the evidence grading system	*
Grade assigned to the **recommendation** with definition of the grade	*
Provide all other grades and definitions from the recommendation grading system	*
Body of evidence: • Quantity – how many studies? • Quality – what type of studies?	*
Estimates of benefit and consistency across studies	*
What harms were identified?	*
Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	*

*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.

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

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

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

1b. Performance Gap

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

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

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

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

N/A

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.*

The measure was calculated using STS Adult Cardiac Surgery Database data for patients undergoing mitral valve repair/replacement (MVRR) +CABG in two consecutive overlapping 3-year time periods, January 2016 – December 2018 and January 2017 – December 2019. For each time period, we provide the number of measured entities (# participants), the number of eligible patient records (# operations), and the distribution of composite score estimates by percentiles and geographic region. Participants with at least 10 eligible records in a 3-year time period were included in the hierarchical model for estimating composite scores in that time period. While participants with 10 eligible cases are included in the hierarchical model procedure, composite scores will typically only be reported by STS for participants with at least 25 cases during a 3-year time period. Thus, we present results for the set of participants with at least 10 eligible cases and the subset with at least 25 eligible cases.

	January 2016-December 2018		January 2017-December 2019	
Distribution	Participants with >=10 Eligible Cases	Participants with >=25 Eligible Cases	Participants with >=10 Eligible Cases	Participants with >=25 Eligible Cases
# Participant	625	289	605	272
# Operations	21383	16175	20403	15087
Mean	0.863	0.866	0.860	0.864
STD	0.02575		0.02745	0.02428	0.02595
IQR	0.0317	0.0352	0.0319	0.0328
0%	0.741	0.741	0.768	0.768
10%	0.830	0.831	0.829	0.831
20%	0.843	0.845	0.841	0.844
30%	0.852	0.854	0.849	0.854
40%	0.859	0.863	0.857	0.861
50%	0.865	0.869	0.862	0.866
60%	0.871	0.875	0.867	0.871
70%	0.877	0.882	0.872	0.878
80%	0.884	0.889	0.879	0.885
90%	0.892	0.897	0.887	0.894
100%	0.936	0.936	0.921	0.921
CANADA	4	2	2	1
MIDWEST	167	66	156	59
NORTHEAST	100	62	99	60
SOUTH	220	115	216	108

WEST 134 44 132 44

If the above table is not clearly displayed, please refer to the version included in the appendix for this measure.

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

N/A

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

In order to shed light on disparities, we used logistic regression to study the associations of race, ethnicity and insurance status with operative mortality and major morbidity while adjusting for covariates included in this measure´s risk adjustment model (see other sections for details of covariate adjustment – we used the most recent 2017 valve + CABG models for mortality and major morbidity). Odds ratios with 95% confidence intervals (CI´s) and p-values are summarized in the table below.

	Mortality		Major Morbidity	
	Adjusted OR (95% CI)	p-value	Adjusted OR (95% CI)	p-value
Insurance status among patients age>=65				
Medicare without Medicaid/Commercial-HMO	(ref)		(ref)	
Medicare + Medicaid dual eligible	0.94(0.71, 1.24)	0.6578	0.81(0.68, 0.98)	0.0287
Medicare + Commercial-HMO without Medicaid	0.97(0.84, 1.13)	0.7131	0.98(0.90, 1.07)	0.6597
Commercial-HMO without Medicare	0.84(0.64, 1.09)	0.1880	1.04(0.88, 1.22)	0.6680
Insurance status among patients age<65				
Commercial-HMO without Medicare/Medicaid	(ref)		(ref)	
Medicare or Medicaid	1.17(0.96, 1.42)	0.1265	1.09(0.98, 1.22)	0.1148
None/Self Paid	0.97(0.65, 1.45)	0.8796	1.02(0.83, 1.25)	0.8393
Other	1.23(0.77, 1.97)	0.3833	1.00(0.76, 1.31)	0.9743
Black Race	0.91(0.75, 1.11)	0.3471	1.28(1.15, 1.43)	<.0001
Hispanic ethnicity	1.13(0.92, 1.39)	0.2510	1.10(0.97, 1.24)	0.1558

If the above table is not clearly displayed, please refer to the version included in the appendix for this measure.

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

1c. Composite Quality Construct and Rationale

1c.1. A composite performance measure is a combination of two or more component measures, each of which individually reflects quality of care, into a single performance measure with a single score.

For purposes of NQF measure submission, evaluation, and endorsement, the following will be considered composites:

- Measures with two or more individual performance measure scores combined into one score for an accountable entity.
- Measures with two or more individual component measures assessed separately for each patient and then aggregated into one score for an accountable entity:
    - all-or-none measures (e.g., all essential care processes received, or outcomes experienced, by each patient);

1c.1. Please identify the composite measure construction: two or more individual performance measure scores combined into one score

1c.2. Describe the quality construct, including:

- the overall area of quality
- included component measures and
- the relationship of the component measures to the overall composite and to each other.

The STS Mitral Valve Repair/Replacement (MVRR) + Coronary Artery Bypass Graft (CABG) Composite Score measures surgical performance for MVRR + CABG with or without concomitant Atrial Septal Defect (ASD) and Patient Foramen Ovale (PFO) closures, tricuspid valve repair (TVr), or surgical ablation for atrial fibrillation (AF). Similar to other STS composite measures, this measure is based on a combination of the NQF-endorsed risk-adjusted operative mortality outcome measure and the risk-adjusted occurrence of any of five major complications. An NQF-endorsed structure measure, database participation, is included de facto as only STS Adult Cardiac Surgery Database participants are eligible to receive composite scores. To assess overall quality, the composite comprises the following two domains:

Domain 1 – Absence of Operative Mortality

Proportion of patients (risk-adjusted) who do not experience operative mortality. Operative mortality is defined as death before hospital discharge or within 30 days of the operation.

Domain 2 – Absence of Major Morbidity

Proportion of patients (risk-adjusted) who do not experience any major morbidity. Major morbidity is defined as the occurrence of any one or more of the following major complications:

1. Prolonged ventilation,
2. Deep sternal wound infection,
3. Permanent stroke,
4. Renal failure, and
5. Reoperations for bleeding, coronary graft occlusion, prosthetic or native valve dysfunction, and other cardiac reasons, but not for other non-cardiac reasons.

Participants receive a score for each of the two domains, plus an overall composite score. The overall composite score was created by “rolling up” the domain scores into a single number. In addition to receiving a numeric score, participants are assigned to rating categories designated by one to three stars.

Similar to the NQF-endorsed STS AVR and AVR+CABG measures, the MVRR+CABG Composite Score differs from the NQF-endorsed STS CABG Composite Score in that it does not include process measures. This reflects

the fact that for MVRR+CABG, in comparison with isolated CABG surgery, no widely accepted process measures meeting performance metric criteria currently exist.

1c.3. Describe the rationale for constructing a composite measure, including how the composite provides a distinctive or additive value over the component measures individually.

Risk-adjusted mortality has historically been the dominant outcomes metric for cardiac surgical procedures. In an era when the average mortality rates for these procedures have declined to very low levels, however, differentiating performance based on mortality alone is inadequate. Specifically, mortality alone fails to account for the fact that not all operative survivors received equal quality care. Patients who survive surgery but have a debilitating complication that may substantially impact their long-term freedom from adverse cardiac events, for example, are not duly reflected in mortality rate measures. A composite score provides a more comprehensive measure of overall surgical quality, and is timely since MVRR+CABG comprises an increasing proportion of cardiac surgical practice, and their mortality risk is higher than for isolated MVRR.

References

- Rankin JS, Badhwar V, He X, Jacobs JP, Gammie JS, Furnary AP, Fazzalari FL, Han J, O'Brien SM, Shahian DM. The Society of Thoracic Surgeons mitral valve repair/replacement plus coronary artery bypass grafting composite score: a report of the Society of Thoracic Surgeons Quality Measurement Task Force. Ann Thorac Surg 2017;103:1475-81.
- Rankin JS, Feneley MP, Hickey M StJ, et al. A clinical comparison of mitral valve repair versus valve replacement in ischemic mitral regurgitation. J Thorac Cardiovasc Surg 1988;95:165-77.
- Glower DD, Tuttle RH, Shaw LK, et al: Patient survival characteristics after routine mitral valve repair for ischemic mitral regurgitation. J Thorac Cardiovasc Surg 2005;129:860-8.
- Milano CA, Daneshmand MA, Rankin JS, et al. Survival prognosis and surgical management of ischemic mitral regurgitation. Ann Thorac Surg 2008;86:735-44.
- Daneshmand MA, Milano CA, Rankin JS, et al. Mitral valve repair for degenerative disease: A 20-year experience. Ann Thorac Surg 2009;88:1828-37.
- Daneshmand MA, Milano CA, Rankin JS, et al. Influence of patient age on procedural selection in mitral valve surgery. Ann Thorac Surg 2010;90:1479-86.

1c.4. Describe how the aggregation and weighting of the component measures are consistent with the stated quality construct and rationale.

The mortality domain corresponds to a single measure, while the study endpoint for the morbidity domain combines multiple measures and thus is a composite endpoint. To enhance interpretation, mortality rates were converted to survival rates (risk-standardized survival rate = 100 – risk-standardized mortality rate), and morbidity rates were converted to "absence of morbidity" rates (risk-standardized absence of morbidity rate =100 – risk-standardized morbidity rate). Defining scores in this manner ensures that increasingly positive values reflect better performance. The overall composite score is created by "rolling up" the domain scores into a single number.

2. Reliability and Validity—Scientific Acceptability of Measure Properties

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

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

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

Cardiovascular, Surgery, Surgery : Cardiac Surgery

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

Safety, Safety : Complications, Safety : Healthcare Associated Infections

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

Adults, Elderly

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

https://www.sts.org/sites/default/files/STSAdultCVDataCollectionFormV4_20_2_GOLDEN006292020.pdf

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)*

No data dictionary **Attachment:**

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

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

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

Not an instrument-based measure

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

No

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

N/A

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

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

Due to the complex methodology used to construct the composite measure, it is impractical to separately discuss the numerator and denominator. The following discussion describes how each domain score is calculated and how these are combined into an overall composite score.

The STS Mitral Valve Repair/Replacement (MVRR) Composite Score comprises two domains consisting of six measures:

Domain 1 – Absence of Operative Mortality

Proportion of patients (risk-adjusted) who do not experience operative mortality. Operative mortality is defined as death before hospital discharge or within 30 days of the operation.

Domain 2 – Absence of Major Morbidity

Proportion of patients (risk-adjusted) who do not experience any major morbidity. Major morbidity is defined as the occurrence of any one or more of the following major complications:

1. Prolonged ventilation,
2. Deep sternal wound infection,
3. Permanent stroke,
4. Renal failure, and
5. Reoperations for bleeding, coronary graft occlusion, prosthetic or native valve dysfunction, and other cardiac reasons, but not for other non-cardiac reasons.

Participants receive a score for each of the two domains, plus an overall composite score. The overall composite score was created by "rolling up" the domain scores into a single number. In addition to receiving a numeric score, participants are assigned to rating categories designated by the following:

1 star – lower-than-expected performance

2 stars – as-expected performance

3 stars – higher-than-expected performance

Patient Population: The analysis population consists of patients aged 18 years or older who MVRR + CABG with or without concomitant Atrial Septal Defect (ASD) and Patient Foramen Ovale (PFO) closures, tricuspid valve repair (TVr), or surgical ablation for atrial fibrillation (AF).

Time Window: 3 years

Data Completeness Requirement: Participants are excluded from the analysis if they have fewer than 25 MVRR + CABG procedures in the patient population.

Estimation of Composite Scores and Star Ratings:

To be consistent with the conventions of previous composite measures, risk-adjusted event rates were first converted into risk-adjusted absence-of-event rates. To calculate the composite, participant-specific absence of mortality rates and absence of morbidity rates were weighted inversely by their respective standard deviations across participants. This procedure was equivalent to first rescaling the absence of mortality rates and absence of morbidity rates by their respective standard deviations across participants, and then assigning equal weighting to the rescaled rates. Finally, in order to draw statistical inferences about participant performance, a Bayesian credible interval surrounding each participant's composite score was calculated. Unlike frequentist confidence intervals, Bayesian credible intervals have an intuitively direct interpretation as an interval containing the true value of the composite score with a specified probability (e.g., 95%). To determine star ratings for each participant, the credible interval of its composite score was compared with the STS average. Participants whose intervals were entirely above the STS average were classified as 3-star (higher than expected performance), and participants whose intervals were entirely below the STS average were classified as 1-star (lower than expected performance). Credible intervals based on different probability levels (90%, 95%, 98%) were explored, and the resulting percentages of 1, 2, and 3-star programs were calculated.

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).

See response in S.4. Numerator Statement

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

See response in S.4. Numerator Statement for complete description of measure specifications.

Patient Population: The analysis population consists of patients aged 18 years or older who MVRR + CABG with or without concomitant Atrial Septal Defect (ASD) and Patient Foramen Ovale (PFO) closures, tricuspid valve repair (TVr), or surgical ablation for atrial fibrillation (AF).See response in S.4. Numerator Statement for complete description of measure specifications.

Patient Population: The analysis population consists of patients aged 18 years or older who MVRR + CABG with or without concomitant Atrial Septal Defect (ASD) and Patient Foramen Ovale (PFO) closures, tricuspid valve repair (TVr), or surgical ablation for atrial fibrillation (AF).

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).

See response in S.7. Denominator Statement

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

Data Completeness Requirement: Participants are excluded from the analysis if they have fewer than 25 MVRR + CABG procedures in the patient population.

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.)*

See response in S.8. Denominator Exclusions

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 = Higher 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.)*

Please see discussion under section S.4 and attached manuscripts.

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

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

N/A

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

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

N/A

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

If other, please describe in S.18.

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.

STS Adult Cardiac Surgery Database – Version 2.73; STS Adult Cardiac Surgery Database Version 2.81 went live on July 1, 2014; STS Adult Cardiac Surgery Database – Version 2.9 went live on July 1st, 2017 and STS Adult Cardiac Surgery Database version 4.20 went live on June 30, 2020.

The URL provided under S.1 is for the latest data collection form that is currently in use.

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

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

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

Clinician : Group/Practice, 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.)*

Please see section S.4

2. Validity – See attached Measure Testing Submission Form

3032_NQF_testing_v3.0-STS_MVRR-CABG_Comp-112320.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.

No

2.2 For maintenance of endorsement

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

No

2.3 For maintenance of endorsement

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

Yes - Updated information is included

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

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

Measure Number (*if previously endorsed*)**:** 3032
Composite Measure Title: STS Mitral Valve Repair/Replacement (MVRR) + CABG Composite Score
Date of Submission: **8/1/2020**

Composite Construction:

☒Two or more individual performance measure scores combined into one score

☐ All-or-none measures (e.g., all essential care processes received or outcomes experienced by each patient)

Instructions: Please contact NQF staff before you begin.

- If a component measure is submitted as an individual performance measure, the non-composite measure testing form must also be completed and attached to the individual measure submission.
- Measures must be tested for all the data sources and levels of analyses that are specified. *If there is more than one set of data specifications or more than one level of analysis, contact NQF staff* about how to present all the testing information in one form.
- Sections 1, 2a2, 2b1, 2b2, and 2b4 must be completed.
- For composites with outcome and resource use measures, section 2b3 also must be completed.
- If specified for multiple data sources/sets of specificaitons (e.g., claims and EHRs), section 2b5 also must be completed.
- Respond to all questions as instructed with answers immediately following the question. All information on testing to demonstrate meeting the subcriteria for reliability (2a2) and validity (2b1-2b6) and composites (2c) must be in this form. An appendix for *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 25 pages (*incuding questions/instructions;* minimum font size 11 pt; do not change margins). *Contact NQF staff if more pages are needed.*
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.
- For information on the most updated guidance on how to address social risk factors variables and testing in this form refer to the release notes for version 7.1 of the Measure Testing Attachment. and the 2017 Measure Evaluation Criteria and Guidance.

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

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

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

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

AND

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

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

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

OR

- rationale/data support no risk adjustment/ stratification.

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

OR

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

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

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

2c. For composite performance measures, empirical analyses support the composite construction approach and demonstrate that:

2c1. the component measures fit the quality construct and add value to the overall composite while achieving the related objective of parsimony to the extent possible; and

2c2. the aggregation and weighting rules are consistent with the quality construct and rationale while achieving the related objective of simplicity to the extent possible.

(*if not conducted or results not adequate, justification must be submitted and accepted)*

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 different components in the composite, indicate the component after the checkbox. If different data sources are used for the numerator and denominator, indicate N [numerator] or D [denominator] after the checkbox.***)

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

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*).

STS Adult Cardiac Surgery Database Version 4.20

1.3. What are the dates of the data used in testing? July 2011 – June 2014

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

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

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

The measure was developed and tested using STS Adult Cardiac Surgery Database data from 703 participants for patients undergoing mitral valve repair/replacement (MVRR) + CABG during July 2011 – June 2014. Only participants with at least 10 eligible records during this period were included in the hierarchical model for estimating composite scores. The table below summarizes the distribution of participant-specific denominators (number of eligible patients) and participant-specific mortality and morbidity rates.

Stat	N (Denominator)	% Mortality	% Morbidity
N	703	703	703
Mean	35	6.8	31.7
STD	34	6.1	13.3

Stat	N (Denominator)	% Mortality	% Morbidity
IQR	27	6.4	17.3
0%	10	0.0	0.0
10%	12	0.0	16.3
20%	14	0.0	20.3
30%	17	3.6	24.0
40%	20	4.8	27.0
50%	23	5.9	30.3
60%	28	7.1	33.3
70%	36	8.6	37.8
80%	50	10.5	42.1
90%	73	14.3	50.0
100%	503	38.9	76.9

Our quality data are collected in the STS National Database at participant-level. As highlighted in the table below, over 92% of STS participants are surgical practice groups that each have a "one to one" relationship with an individual hospital. Therefore, with the exception of measures specifically identified as individual surgeon-focused (currently only the STS Individual Surgeon Composite for Adult Cardiac Surgery, NQF# 3030), STS performance measures are developed and validated at the STS participant level and do not require multiple levels of analysis.

Please note that the data in the table below includes all participants in the Adult Cardiac Surgery Database (ACSD) and is not specific to the subset of ACSD participants for whom data are reported for this specific measure.

Distribution of STS "Participant" Contract Types in Adult Cardiac Surgery Database (11/2/2020)	**# of Participants**	**% of Participants**
Surgeon group only without hospital (including groups providing services at multiple hospitals), i.e., one-to-many	31	3.00%
Surgeon group w/individual hospital, i.e., one-to-one	952	92.40%
Surgeon group w/no hospital listed, i.e., new participant still being set up	2	0.20%
Individual surgeon	45	4.40%
Total US & Canada Participants	1030	100%

There is considerable sample size variation within and across different STS "participant" categories. To assure that our methodology is valid and reliable for any "participant" to whom we provide a score, we conduct sophisticated Markov Chain Monte Carlo (MCMC) simulations for all our measures to test their average reliability at different volume thresholds. STS estimates minimum sample size (i.e., case volume) thresholds, with their corresponding reliabilities, for each measure. We require that any participant receiving an STS score must have a volume of cases of the specific case type, during the prescribed analytic timeframe (i.e.,

typically 1 or 3 years), that assures an average reliability of 0.50, one of the highest measure reliability standards of which we are aware in all of healthcare.

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*)
For assessing risk model discrimination and calibration, the sample included eligible 26,355 patient operation records from 1,038 STS participants. For estimating composite scores, the sample was limited to participants with at least 10 eligible cases over the 3-year study period (24,740 patients, 703 centers).

The table below summarizes the baseline characteristics of patients who were included in the estimation of composite scores during July 2011 – June 2014.

Variable	Effects	Overall N=24,740	MV Repair N=16,300	MV Replace N=8,440
Patient Age	Median (IQR)	69.0 (62.0, 76.0)	69.0 (61.0, 76.0)	70.0 (62.0, 77.0)
Surgery Year	2011 (half year)	3,958 (16.0%)	2,664 (16.3%)	1,294 (15.3%)
*	2012	8,230 (33.3%)	5,525 (33.9%)	2,705 (32.0%)
*	2013	8,417 (34.0%)	5,571 (34.2%)	2,846 (33.7%)
*	2014 (half year)	4,135 (16.7%)	2,540 (15.6%)	1,595 (18.9%)
Previous Cardiac Surgery	No	16,167 (65.3%)	11,047 (67.8%)	5,120 (60.7%)
*	Yes	8,532 (34.5%)	5,228 (32.1%)	3,304 (39.1%)
*	Missing	41 (0.2%)	25 (0.2%)	16 (0.2%)
Previous PCI	No	2,411 (28.3%)	1,240 (23.7%)	1,171 (35.4%)
*	Yes	6,112 (71.6%)	3,987 (76.3%)	2,125 (64.3%)
*	Missing	9 (0.1%)	1 (0.0%)	8 (0.2%)
Previous PCI - Interval	<=6 Hours	88 (1.4%)	39 (1.0%)	49 (2.3%)
*	>6 Hours	6,009 (98.3%)	3,939 (98.8%)	2,070 (97.4%)
*	Missing	15 (0.2%)	9 (0.2%)	6 (0.3%)
Ejection Fraction (%)	Median (IQR)	48.0 (35.0, 60.0)	45.0 (30.0, 56.0)	53.0 (40.0, 60.0)
*	Missing	428 (1.7%)	255 (1.6%)	173 (2.0%)
LV End-Systolic Dimension	Median (IQR)	38.0 (31.3, 46.0)	39.0 (33.0, 47.2)	35.9 (29.8, 43.0)
*	Missing	12,082 (48.8%)	7,798 (47.8%)	4,284 (50.8%)
LV End-Diastolic Dimension	Median (IQR)	53.0 (47.0, 59.0)	54.0 (48.0, 59.0)	51.0 (45.2, 57.0)
*	Missing	12,060 (48.7%)	7,769 (47.7%)	4,291 (50.8%)
PA Systolic Pressure	Median (IQR)	43.0 (33.8, 55.0)	42.0 (32.0, 53.0)	46.0 (35.0, 59.0)
*	Missing	36 (0.3%)	24 (0.3%)	12 (0.2%)
Aortic Valve Disease (mild)	No	17,294 (69.9%)	11,554 (70.9%)	5,740 (68.0%)
*	Yes	7,340 (29.7%)	4,673 (28.7%)	2,667 (31.6%)
*	Missing	106 (0.4%)	73 (0.4%)	33 (0.4%)
Aortic Valve Disease Etiol.	Degenerative	2,536 (34.6%)	1,544 (33.0%)	992 (37.2%)
*	Endocarditis	19 (0.3%)	9 (0.2%)	10 (0.4%)
*	Congenital	32 (0.4%)	24 (0.5%)	8 (0.3%)
*	Rheumatic	74 (1.0%)	13 (0.3%)	61 (2.3%)
*	Primary Aortic Dis.	12 (0.2%)	6 (0.1%)	6 (0.2%)
*	LVOTO	8 (0.1%)	2 (0.0%)	6 (0.2%)
*	Tumor	2 (0.0%)	1 (0.0%)	1 (0.0%)
*	Trauma	1 (0.0%)	1 (0.0%)	0 (0.0%)

Variable	Effects	Overall N=24,740	MV Repair N=16,300	MV Replace N=8,440
*	Other	1,080 (14.7%)	707 (15.1%)	373 (14.0%)
*	Missing	3,574 (48.7%)	2,366 (50.6%)	1,208 (45.3%)
Aortic Valve Stenosis	No	6,269 (85.4%)	4,090 (87.5%)	2,179 (81.7%)
*	Yes	909 (12.4%)	465 (10.0%)	444 (16.6%)
*	Missing	162 (2.2%)	118 (2.5%)	44 (1.6%)
Aortic Valve Insufficiency	None	541 (7.4%)	316 (6.8%)	225 (8.4%)
*	Trivial	2,884 (39.3%)	1,917 (41.0%)	967 (36.3%)
*	Mild	3,116 (42.5%)	1,964 (42.0%)	1,152 (43.2%)
*	Moderate	742 (10.1%)	440 (9.4%)	302 (11.3%)
*	Severe	40 (0.5%)	26 (0.6%)	14 (0.5%)
*	Missing	17 (0.2%)	10 (0.2%)	7 (0.3%)
MV Disease Etiol.	Degenerative	12,807 (53.0%)	8,877 (55.9%)	3,930 (47.3%)
*	Endocarditis	643 (2.7%)	174 (1.1%)	469 (5.6%)
*	Rheumatic	1,129 (4.7%)	147 (0.9%)	982 (11.8%)
*	Ischemic	3,300 (13.6%)	2,357 (14.8%)	943 (11.3%)
*	Congenital	73 (0.3%)	56 (0.4%)	17 (0.2%)
*	HOCM	39 (0.2%)	20 (0.1%)	19 (0.2%)
*	Tumor	77 (0.3%)	49 (0.3%)	28 (0.3%)
*	Trauma	5 (0.0%)	4 (0.0%)	1 (0.0%)
*	Non-isch Myopathy	153 (0.6%)	103 (0.6%)	50 (0.6%)
*	Other	2,028 (8.4%)	1,287 (8.1%)	741 (8.9%)
*	Missing	3,930 (16.3%)	2,799 (17.6%)	1,131 (13.6%)
MV Degenerative Location	Posterior Leaflet	953 (34.7%)	780 (41.1%)	173 (20.4%)
*	Anterior Leaflet	303 (11.0%)	178 (9.4%)	125 (14.7%)
*	Bi-leaflet	445 (16.2%)	225 (11.9%)	220 (25.9%)
*	Missing	1,043 (38.0%)	713 (37.6%)	330 (38.9%)
Mitral Annular Disease Type	Pure Ann Dilation	950 (39.7%)	872 (47.8%)	78 (13.8%)
*	Ann Calcification	397 (16.6%)	238 (13.1%)	159 (28.0%)
*	Missing	1,043 (43.6%)	713 (39.1%)	330 (58.2%)
MV Ischemic Type	Acute	1,105 (33.5%)	638 (27.1%)	467 (49.5%)
*	Chronic	2,081 (63.1%)	1,630 (69.2%)	451 (47.8%)
*	Missing	114 (3.5%)	89 (3.8%)	25 (2.7%)
MV NYHA Functional Class	I	4,528 (18.7%)	3,403 (21.4%)	1,125 (13.5%)
*	II	5,429 (22.4%)	3,612 (22.8%)	1,817 (21.9%)
*	IIIa	2,926 (12.1%)	1,509 (9.5%)	1,417 (17.0%)
*	IIIb	2,222 (9.2%)	1,476 (9.3%)	746 (9.0%)
*	Missing	9,079 (37.5%)	5,873 (37.0%)	3,206 (38.6%)
Mitral Stenosis	No	21,815 (90.2%)	15,174 (95.6%)	6,641 (79.9%)
*	Yes	1,775 (7.3%)	274 (1.7%)	1,501 (18.1%)
*	Missing	594 (2.5%)	425 (2.7%)	169 (2.0%)
Mitral Insufficiency	None	260 (1.1%)	87 (0.5%)	173 (2.1%)
*	Trivial	242 (1.0%)	118 (0.7%)	124 (1.5%)
*	Mild	1,070 (4.4%)	625 (3.9%)	445 (5.4%)

Variable	Effects	Overall N=24,740	MV Repair N=16,300	MV Replace N=8,440
*	Moderate	5,722 (23.7%)	4,374 (27.6%)	1,348 (16.2%)
*	Severe	16,795 (69.4%)	10,606 (66.8%)	6,189 (74.5%)
*	Missing	95 (0.4%)	63 (0.4%)	32 (0.4%)
Tricuspid Valve Disease	No	9,480 (38.3%)	6,266 (38.4%)	3,214 (38.1%)
*	Yes	15,155 (61.3%)	9,961 (61.1%)	5,194 (61.5%)
*	Missing	105 (0.4%)	73 (0.4%)	32 (0.4%)
Tricuspid Insufficiency	None	57 (0.4%)	36 (0.4%)	21 (0.4%)
*	Trivial	3,159 (20.8%)	2,254 (22.6%)	905 (17.4%)
*	Mild	6,052 (39.9%)	4,020 (40.4%)	2,032 (39.1%)
*	Moderate	4,137 (27.3%)	2,603 (26.1%)	1,534 (29.5%)
*	Severe	1,671 (11.0%)	1,000 (10.0%)	671 (12.9%)
*	Missing	79 (0.5%)	48 (0.5%)	31 (0.6%)
*	Missing	8 (0.1%)	4 (0.1%)	4 (0.2%)
Operative Approach	Full Sternotomy	24,560 (99.3%)	16,191 (99.3%)	8,369 (99.2%)
*	Partial Sternotomy	71 (0.3%)	41 (0.3%)	30 (0.4%)
*	Rt or Lt Parasternal	4 (0.0%)	2 (0.0%)	2 (0.0%)
*	Left Thoracotomy	6 (0.0%)	1 (0.0%)	5 (0.1%)
*	Right Thoracotomy	14 (0.1%)	7 (0.0%)	7 (0.1%)
*	Transverse	1 (0.0%)	1 (0.0%)	0 (0.0%)
*	Minimally Invasive	27 (0.1%)	24 (0.1%)	3 (0.0%)
*	Missing	57 (0.2%)	33 (0.2%)	24 (0.3%)
Robotic Assisted	No	24,472 (98.9%)	16,116 (98.9%)	8,356 (99.0%)
*	Yes	39 (0.2%)	26 (0.2%)	13 (0.2%)
*	Missing	229 (0.9%)	158 (1.0%)	71 (0.8%)
Mitral Valve Procedure	Repair	16,300 (65.9%)	16,300 (100.0%)	0 (0.0%)
*	Replacement	8,440 (34.1%)	0 (0.0%)	8,440 (100.0%)
MV Repair - Annuloplasty	No	729 (4.5%)	729 (4.5%)	. (.%)
*	Yes	15,434 (94.7%)	15,434 (94.7%)	. (.%)
*	Missing	137 (0.8%)	137 (0.8%)	. (.%)
MV Repair - Leaf Resection	No	13,241 (81.2%)	13,241 (81.2%)	. (.%)
*	Yes	2,660 (16.3%)	2,660 (16.3%)	. (.%)
*	Missing	399 (2.4%)	399 (2.4%)	. (.%)
MV Leaflet Resection Type	Triangular	1,323 (49.7%)	1,323 (49.7%)	. (.%)
*	Quadrangular	991 (37.3%)	991 (37.3%)	. (.%)
*	Other	255 (9.6%)	255 (9.6%)	. (.%)
*	Missing	91 (3.4%)	91 (3.4%)	. (.%)
MV Repair Location	Anterior	181 (6.8%)	181 (6.8%)	. (.%)
*	Posterior	2,279 (85.7%)	2,279 (85.7%)	. (.%)
*	Bileaflet	143 (5.4%)	143 (5.4%)	. (.%)
*	Missing	57 (2.1%)	57 (2.1%)	. (.%)
MV Sliding Plasty	No	15,290 (93.8%)	15,290 (93.8%)	. (.%)
*	Yes	557 (3.4%)	557 (3.4%)	. (.%)
*	Missing	453 (2.8%)	453 (2.8%)	. (.%)

Variable	Effects	Overall N=24,740	MV Repair N=16,300	MV Replace N=8,440
MV Annular Decalcification	No	15,740 (96.6%)	15,740 (96.6%)	. (.%)
*	Yes	136 (0.8%)	136 (0.8%)	. (.%)
*	Missing	424 (2.6%)	424 (2.6%)	. (.%)
PTFE Chordal Replacement	No	14,529 (89.1%)	14,529 (89.1%)	. (.%)
*	Yes	1,316 (8.1%)	1,316 (8.1%)	. (.%)
*	Missing	455 (2.8%)	455 (2.8%)	. (.%)
Neo-chordal Number	Median (IQR)	2.0 (1.0, 3.0)	2.0 (1.0, 3.0)	. (., .)
*	Missing	39 (3.0%)	39 (3.0%)	. (.%)
MV Chordal Transfer	0	15,594 (95.7%)	15,594 (95.7%)	. (.%)
*	Yes	263 (1.6%)	263 (1.6%)	. (.%)
*	Missing	443 (2.7%)	443 (2.7%)	. (.%)
MV Leaflet Patch	0	15,641 (96.0%)	15,641 (96.0%)	. (.%)
*	Yes	215 (1.3%)	215 (1.3%)	. (.%)
*	Missing	444 (2.7%)	444 (2.7%)	. (.%)
MV Edge To Edge Repair	No	15,189 (93.2%)	15,189 (93.2%)	. (.%)
*	Yes	640 (3.9%)	640 (3.9%)	. (.%)
*	Missing	471 (2.9%)	471 (2.9%)	. (.%)
Mitral Commissurotomy	0	667 (4.1%)	667 (4.1%)	. (.%)
*	Yes	15,633 (95.9%)	15,633 (95.9%)	. (.%)
MV Repair Attempt	No	6,989 (82.8%)	. (.%)	6,989 (82.8%)
*	Yes	1,260 (14.9%)	. (.%)	1,260 (14.9%)
*	Missing	191 (2.3%)	. (.%)	191 (2.3%)
MV Replace - Chordal Pres.	None	3,330 (13.5%)	1,651 (10.1%)	1,679 (19.9%)
*	Anterior	607 (2.5%)	344 (2.1%)	263 (3.1%)
*	Posterior	1,938 (7.8%)	186 (1.1%)	1,752 (20.8%)
*	Both	12,512 (50.6%)	8,865 (54.4%)	3,647 (43.2%)
*	Missing	6,353 (25.7%)	5,254 (32.2%)	1,099 (13.0%)
*	Missing	847 (3.4%)	549 (3.4%)	298 (3.5%)
Postop TEE MR Grade	None	10,633 (55.2%)	6,683 (51.2%)	3,950 (63.6%)
*	Trace/Trivial	4,718 (24.5%)	3,548 (27.2%)	1,170 (18.8%)
*	Mild	1,653 (8.6%)	1,399 (10.7%)	254 (4.1%)
*	Moderate	682 (3.5%)	538 (4.1%)	144 (2.3%)
*	Severe	662 (3.4%)	395 (3.0%)	267 (4.3%)
*	Missing	918 (4.8%)	494 (3.8%)	424 (6.8%)
Operative Mortality	No	23,208 (93.8%)	15,500 (95.1%)	7,708 (91.3%)
*	Yes	1,532 (6.2%)	800 (4.9%)	732 (8.7%)
Any Post-Op Events	No	9,213 (37.2%)	6,531 (40.1%)	2,682 (31.8%)
*	Yes	15,475 (62.6%)	9,736 (59.7%)	5,739 (68.0%)
*	Missing	52 (0.2%)	33 (0.2%)	19 (0.2%)
Reop - Bleeding	No	23,683 (95.7%)	15,679 (96.2%)	8,004 (94.8%)
*	Yes	984 (4.0%)	575 (3.5%)	409 (4.8%)
*	Missing	73 (0.3%)	46 (0.3%)	27 (0.3%)
Reop - Valve Dysfunction	No	24,621 (99.5%)	16,222 (99.5%)	8,399 (99.5%)

Variable	Effects	Overall N=24,740	MV Repair N=16,300	MV Replace N=8,440
*	Yes	44 (0.2%)	30 (0.2%)	14 (0.2%)
*	Missing	75 (0.3%)	48 (0.3%)	27 (0.3%)
Reop - Other Cardiac	No	24,300 (98.2%)	16,028 (98.3%)	8,272 (98.0%)
*	Yes	364 (1.5%)	225 (1.4%)	139 (1.6%)
*	Missing	76 (0.3%)	47 (0.3%)	29 (0.3%)
Reop - Other Non-Cardiac	No	23,570 (95.3%)	15,612 (95.8%)	7,958 (94.3%)
*	Yes	1,094 (4.4%)	639 (3.9%)	455 (5.4%)
*	Missing	76 (0.3%)	49 (0.3%)	27 (0.3%)
Deep Sternal Infection	No	24,559 (99.3%)	16,191 (99.3%)	8,368 (99.1%)
*	Yes	102 (0.4%)	60 (0.4%)	42 (0.5%)
*	Missing	79 (0.3%)	49 (0.3%)	30 (0.4%)
Permanent Stroke	No	23,975 (96.9%)	15,824 (97.1%)	8,151 (96.6%)
*	Yes	684 (2.8%)	423 (2.6%)	261 (3.1%)
*	Missing	81 (0.3%)	53 (0.3%)	28 (0.3%)
Prolonged Ventilation	No	18,216 (73.6%)	12,499 (76.7%)	5,717 (67.7%)
*	Yes	6,428 (26.0%)	3,731 (22.9%)	2,697 (32.0%)
*	Missing	96 (0.4%)	70 (0.4%)	26 (0.3%)
Renal Failure	No	23,021 (93.1%)	15,320 (94.0%)	7,701 (91.2%)
*	Yes	1,647 (6.7%)	934 (5.7%)	713 (8.4%)
*	Missing	72 (0.3%)	46 (0.3%)	26 (0.3%)

*cell intentionally left blank

PCI=percutaneous coronary intervention, LV=left ventricular, PA=pulmonary artery, Etiol.=etiology, MV=mitral valve, NYHA=New York Heart Association, PTFE=polytetrafluoroethylene, Pres=preservation, TEE=transesophageal echo, Reop=reoperation

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.

For developing and evaluating case mix adjustment procedures, we used data from 26,355 eligible patients at 1,038 participants undergoing MVRR+CABG during July 2011 – June 2014.

For estimating participant-specific composite scores, the analysis was restricted to data from STS participants with at least 10 eligible cases during July 2011 – June 2014 (N = 24,740 patient records, 703 participants).

For assessing the consistency of results over time, we re-estimated composite scores using data from July 2012 – June 2015 (N = 24,376 patient records, 688 participants). This analysis included all participants with at least 10 eligible cases during July 2012 – June 2015.

To ensure adequate statistical precision, the STS plans to report composite scores only for participants with at least 25 eligible cases during the 3-year measurement window. Thus, some of the analyses in this submission are limited to participants with at least 25 eligible cases.

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.

The STS position on inclusion of social risk factors (e.g., SES/SDS/race) as risk model variables is best summarized in this excerpt from our 2018 risk model publication [1]. We describe in detail the controversies about such variables, and how we have attempted to reconcile them:

> *"Whether outcomes measures, and the public reporting and reimbursement programs based on them, should consider socioeconomic status (SES) or sociodemographic factors (e.g., race, ethnicity, education, income, payer [e.g., Medicare-Medicaid dual eligible status]) is a topic of intense health policy debate. Some argue that in the absence of adjustment for these variables, the outcomes of hospitals that care for a disproportionate percentage of low SES patients will be unfairly disadvantaged, perhaps leading to financial or reputational penalties. Opponents argue that inclusion of SES factors in risk models may "adjust away" disparities in quality of care, and they advocate the use of stratified analyses instead. Also, readily available SES factors have often not demonstrated significant impact on outcomes, perhaps because they are not sufficiently granular or relevant. Finally, even SES proponents agree that these factors make more sense conceptually for some outcomes (e.g., readmission) than for others (hospital mortality, complications). Notably, as part of a National Quality Forum pilot project, the STS specifically studied dual eligible status in the STS readmission measure and found minimal impact. In developing the new STS risk models, we avoided these more philosophical and health policy arguments regarding SES adjustment and based our modeling decisions on empiric findings and consideration of the model's primary intended purpose—optimal case mix adjustment. Conceptually, our goal was to adjust for all preoperative factors that are independently and significantly associated with outcomes and that vary across STS participants. For example, race will continue to be in our risk models as it has been previously, but not conceptually as a SES indicator [Note: nor as a surrogate for such factors]. Race has an empiric association with outcomes and has the potential to confound the interpretation of a hospital's outcomes, although we do not know the underlying mechanism (e.g., genetic factors, differential effectiveness of certain medications, rates of certain associated diseases such as diabetes and hypertension, and potentially SES for some outcomes such as readmission)."*

STS is aware of the recent NEJM paper by Vyas and colleagues [2] and has directly communicated with the lead author to explain why race is included in STS models, and to correct several misinterpretations and misrepresentations in this article. Dr. Vyas acknowledged that they included extended quotes from our risk model paper precisely because we were one of the few risk model developers that thoroughly described our rationale for race inclusion, as noted in the excerpt above.

Documents produced by NQF [3, 4], the National Academy of Medicine [5-8], the Office of the Assistant Secretary for Planning and Evaluation (*Social Risk Factors and Performance Under Medicare's Value-Based Purchasing Programs*) [9], and as part of the 21st Century Cures Act legislation [10] are particularly instructive. They summarize the arguments for and against inclusion of SDS/SES/racial adjustment in risk models; context-specific considerations for when they might be appropriate or inappropriate; strategies to avoid the potential adverse unintended consequences of such adjustment; concomitant monitoring for social and racial inequities through stratification; and special approaches for providers who care for high proportions of disadvantaged populations (e.g., payment adjustments, additional resources).

Adjustment for SDS/SES/racial factors has generally been regarded as acceptable when there is both an *empirical association* AND a *plausible conceptual association* of the risk variable with an outcome. For example, an SES/SDS/racial risk factor might be appropriate as a risk variable for readmission or mortality risk models, but not for CAUTI (catheter-associated urinary tract infections), CLABSI (central line-associated bloodstream infection), or process measures.

For many outcomes, SES/SDS/racial adjustment is warranted to optimize risk model accuracy. For example, STS and Duke Clinical Research Institute analyses show that if race variables are excluded from some STS models, the resulting outcomes estimates are markedly different than the actual observed outcomes, and the O/E ratios are significantly different than unity, especially when the models are applied to racial minority subpopulations. Use of risk estimates from such models for patient counseling and shared decision-making would be misleading to patients and would inaccurately portray the risk-adjusted performance of providers, especially those caring for minority populations.

Although SDS/SES/racial risk adjustment may be indicated to assure optimal risk model estimates based on current data, it could potentially obscure disparities in care. To avoid this unintended consequence, most of the guidance documents cited above recommend that any risk model results that are adjusted for SES/SDS/racial factors also present concomitant results in which outcomes are *stratified by the same variables*. This is a much more direct and explicit approach to monitor disparities and inequities and has been followed by STS in its risk modeling and performance measures. Please refer to the race-specific disparities data provided for each of the domains (mortality and morbidity) of measure 3032 under question 1b.4 (Importance tab) of the submission form ***(to be completed by the November submission deadline)***, which we believe will suffice to comply with this recommendation.

1. Shahian DM, Jacobs JP, Badhwar V, Kurlansky PA, Furnary AP, Cleveland JC, Jr., et al. The Society of Thoracic Surgeons 2018 Adult Cardiac Surgery Risk Models: Part 1-Background, Design Considerations, and Model Development. Ann Thorac Surg. 2018;105(5):1411-8.

2. Vyas DA, Eisenstein LG, Jones DS. Hidden in Plain Sight — Reconsidering the Use of Race Correction in Clinical Algorithms. New England Journal of Medicine. 2020.

3. National Quality Forum. Risk adjustment for Socioeconomic Status or other Sociodemographic Factors, accessed at http://www.qualityforum.org/Publications/2014/08/Risk_Adjustment_for_Socioeconomic_Status_or_Other_Sociodemographic_Factors.aspx on June 24, 2020. 2014.

4. The National Quality Forum. Evaluation of the NQF Trial Period for Risk Adjustment for Social Risk Factors. January 15, 2017. Available from: https://www.qualityforum.org/Publications/2017/07/Social_Risk_Trial_Final_Report.aspx.

5. National Academies of Sciences, Engineering, and Medicine. Accounting for social risk factors in Medicare payment. Washington, DC: The National Academies Press; 2017.

6. National Academies of Sciences, Engineering, and Medicine. Accounting for social risk factors in Medicare payment: Data. Washington, DC; 2016.

7. National Academies of Sciences, Engineering, Medicine. Accounting for Social Risk Factors in Medicare Payment: Criteria, Factors, and Methods. Washington, DC: The National Academies Press; 2016.

8. National Academies of Sciences, Engineering, Medicine,. Accounting for Social Risk Factors in Medicare Payment: Identifying Social Risk Factors. Washington, DC: The National Academies Press; 2016. 110 p.

9. Office of the Assistant Secretary for Planning and Evaluation USDoHaHS. Report to Congress: Social Risk Factors and Performance Under Medicare's Value-Based Purchasing Programs. A Report Required by the Improving Medicare Post-Acute Care Transformation (IMPACT) Act of 2014. Washington, DC; 2016.

10. 114th Congress of the United States. 21st Century Cures Act (Public Law 114–255). Washington, DC; 2016.

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*)

***Note**: Current guidance for composite measure evaluation states that reliability must be demonstrated for the composite performance measure score.*

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

2a2.2. 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*)

Reliability is conventionally defined as the proportion of variation in a performance measure that is due to true between-hospital differences (i.e., signal) as opposed to random statistical fluctuations (i.e., noise). A mathematically equivalent definition is the squared correlation between a measurement and the true value. This quantity cannot be calculated directly because the "true" composite measure values are unknown, but may be estimated, as described below.

Calculation Details

Let θ denote the true unknown composite measure value for the j-th of J participants. Before estimating reliability, the numeric value of θ_j was estimated for each participant under the assumed hierarchical model. Estimation was done using Markov Chain Monte Carlo (MCMC) simulations and involved the following steps:

1. For each *j,* we randomly generated a large number (*N*) of possible numeric values of θ by sampling from the Bayesian posterior probability distribution of θ_j via MCMC sampling. Let $\theta_j^{(i)}$ denote the i-th of these *N* randomly sampled numerical values for the j-th participant.
2. For each j, the posterior mean $\hat{\theta}_j$ of θ_j was calculated as the arithmetic average of the randomly sampled values $\theta_j^{(1)}, \ldots, \theta_j^{(N)}$; in other words $\hat{\theta}_j = \frac{1}{N}\sum_{i=1}^{N} \theta_j^{(i)}$.

Our reliability measure was defined as the squared correlation between the set of hospital-specific estimates $\hat{\theta}_1, \ldots, \hat{\theta}_J$ and the corresponding unknown true values $\theta_1, \ldots, \theta_J$. Let ρ^2 denote the unknown true squared correlation of interest and let $\hat{\rho}^2$ denote an estimate of this quantity. The estimate was calculated as

$$\hat{\rho}^2 = \frac{1}{N}\sum_{i=1}^{N} \rho_{(i)}^2$$

where

$$\rho_{(i)}^2 = \frac{\left[\sum_{j=1}^{J}\left(\theta_j^{(i)} - \bar{\theta}^{(i)}\right)\left(\hat{\theta}_j - \bar{\theta}\right)\right]^2}{\sum_{j=1}^{J}\left(\theta_j^{(i)} - \bar{\theta}^{(i)}\right)^2 \sum_{j=1}^{J}\left(\hat{\theta}_j - \bar{\theta}\right)^2}, \quad \bar{\theta} = \frac{1}{JN}\sum_{j=1}^{J}\sum_{i=1}^{N} \theta_j^{(i)} \quad \text{and} \quad \bar{\theta}^{(i)} = \frac{1}{J}\sum_{j=1}^{J} \theta_j^{(i)}.$$

A 95% Bayesian probability interval for ρ^2 was obtained calculating the 2.5th and 97.5th percentiles of the set of numbers $\rho_{(1)}^2, \ldots, \rho_{(N)}^2$.

2a2.3. 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*)

The estimated reliability of the STS MVRR+CABG composite measure using 3 years of data in participants with at least 25 total cases was 0.50 (95% CrI, 0.44 to 0.57), as outlined in the table below. For comparison, the

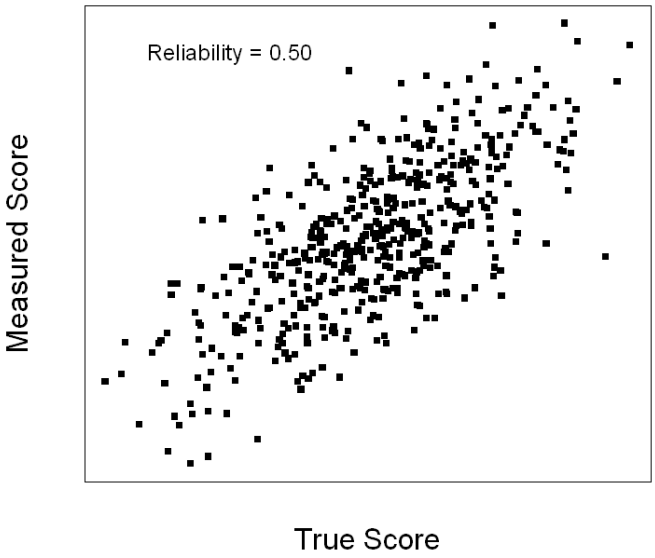
reliability of the STS isolated CABG composite score was 0.77 (95% CrI, 0.74 to 0.80) using 1 year of data in 2013. Using 3 years of data from 2011 to 2013, the reliability of the STS AVR composite measure was 0.52 (95% CrI, 0.47 to 0.57), and the AVR+CABG measure was 0.50 (95% CrI, 0.45 to 0.54)

Time Span	Number of Participants Included	Number of Patients Included	Reliability $\hat{\rho}^2$(95% PrI)
3 years	703	24740	0.42 (0.35, 0.48)
3 years, participants with at least 25 cases	341	18924	0.50 (0.44, 0.57)
3 years, participants with at least 50 cases	143	12217	0.62 (0.52, 0.70)

Based in part on these results, we selected a threshold of 25 cases over 3 years, as a minimum threshold for receiving a site-specific STS MVRR+CABG composite score. This resulted in a reliability of 0.50 but reduced the number of programs eligible to receive a score from 703 to 341. A higher volume threshold would have yielded even higher reliability but at the cost of further reducing the number of programs eligible to receive a score.

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?*)

To interpret the results, we created a figure illustrating the accuracy of the measured scores when the true reliability = 0.50. Because the true score for the composite measure is unknown, we used simulated data with formula Measured Score$_i$=True Score$_i$ + e_i where $i = 1,2, \ldots ,341$ indicates the 341 participants and where True Score$_i$ and e_i both follow normal distributions. The standard deviations of the normal distributions were chosen such that the measure has a reliability of 0.50.



2b1. VALIDITY TESTING

Note*: Current guidance for composite measure evaluation states that validity should be demonstrated for the composite performance measure score. If not feasible for initial endorsement, acceptable alternatives include assessment of content or face validity of the composite OR demonstration of validity for each component. Empirical validity testing of the composite measure score is expected by the time of endorsement maintenance.*

2b1.1. What level of validity testing was conducted?

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

☒ **Composite 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.

☐ **Validity testing for component measures** *(check all that apply)*

***Note**: applies to ALL component measures, unless already endorsed or are being submitted for individual endorsement.*

☐ **Endorsed (or submitted) as individual performance measures**

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

☐ **Empirical validity testing of the component measure score(s)**

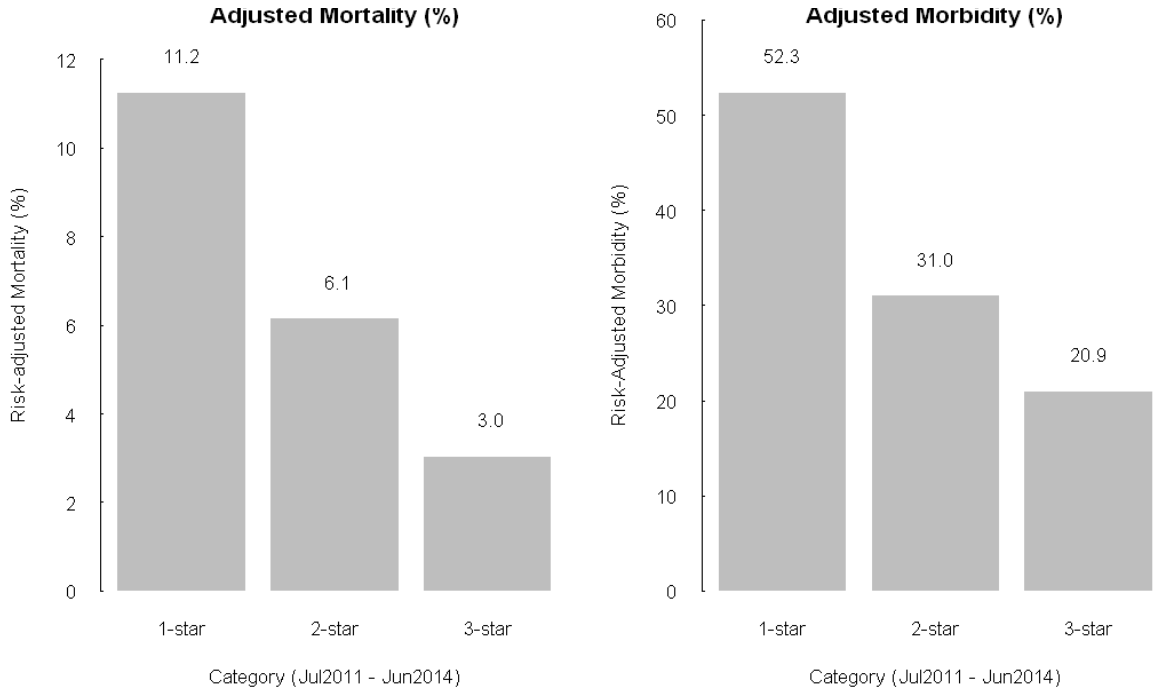
☐ **Systematic assessment of face validity of component measure score(s) as an indicator** of quality or resource use (*i.e., is an accurate reflection of performance on quality or resource use and can distinguish good from poor performance*)

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)*
The tests on validity used the concept of performance categories to be more formally introduced in 2b4: Participants were labeled as having higher-than-expected performance if the 95% credible interval surrounding a participant's composite score fell entirely above the overall STS average composite score. Participants were labeled as having lower-than-expected performance if the 95% credible interval surrounding a participant's composite score fell entirely below the overall STS average composite score. Participants were labeled as higher-than-expected performance (3 stars), lower-than-expected performance (1 star), and indistinguishable from the average or as-expected performance (2 stars).

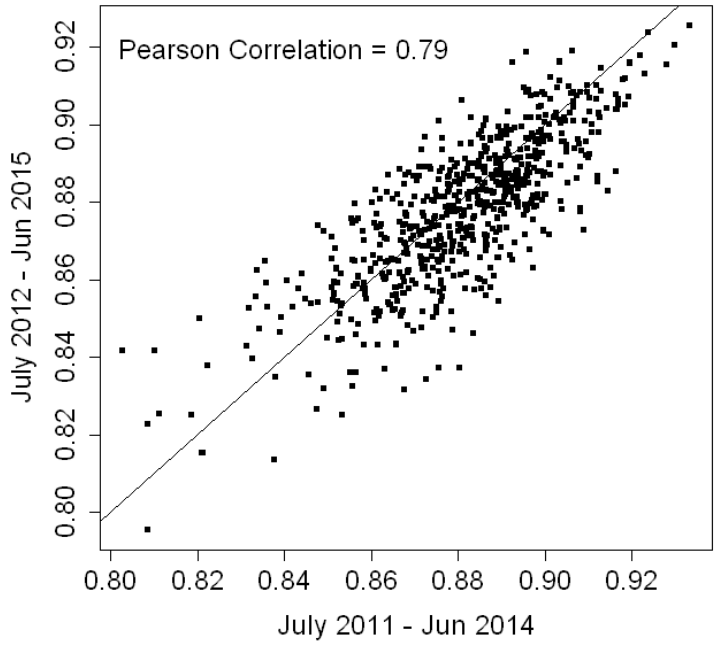
We compared risk-adjusted mortality and morbidity rates across the three performance groups. The measure has good face value if the three groups have different proportions as expected.

In addition, we assessed the extent to which a participant's composite score remains stable across two consecutive overlapping reporting periods. This analysis was restricted to 654 participants who participated and had at least 10 eligible cases in each of two consecutive reporting periods: July 2011 – June 2014 and July 2012 – June 2015.

2b1.3. What were the statistical results from validity testing? (*e.g., correlation; t-test*)
Compared to participants receiving 1 star, those with 3 stars had lower risk-adjusted mortality (3.0% vs. 11.2%) and lower risk-adjusted morbidity (20.9% vs. 52.3%) during July 2011 – June 2014. Thus, differences in performance were clinically meaningful as well as statistically significant. STS participants deemed better by the composite scores have (on average) higher performance during the same time window on each individual domain of the composite measure.

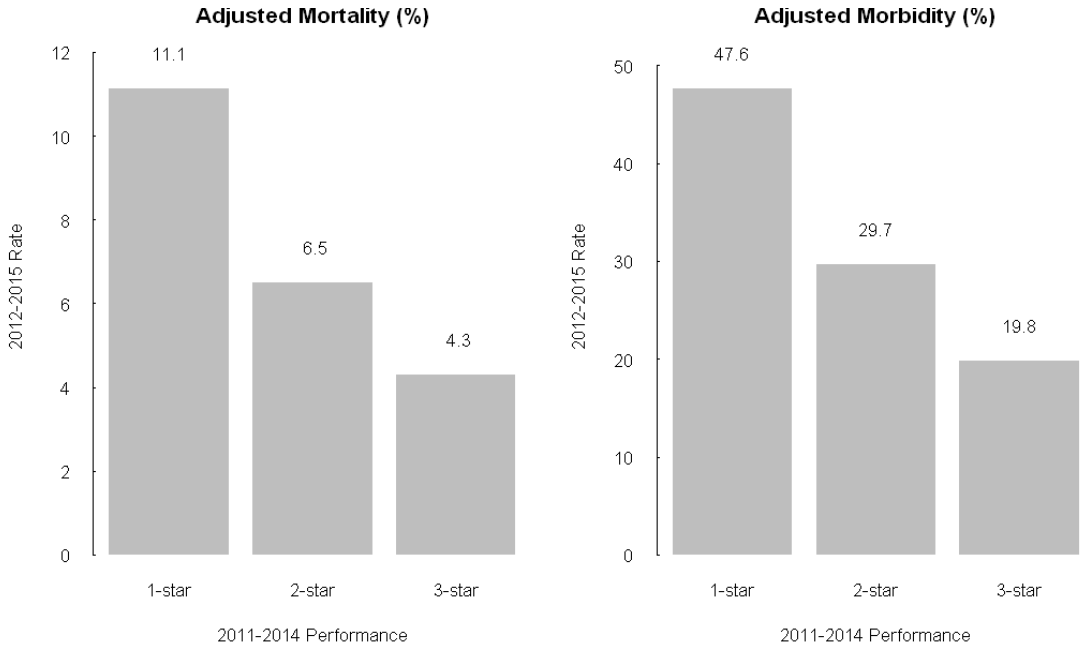


Stability of the composite measure over time was assessed in 654 participants who participated and had at least 10 eligible cases in each of two consecutive reporting periods: July 2011 – June 2014 and July 2012 – June 2015.



The Pearson correlation between the composite score calculated in the earlier and later time period was 0.79.

Using data from July 2012 – June 2015, we compared risk-adjusted mortality and morbidity rates across participants categories based on their composite measure performance in July 2011 – June 2014. Compared to 1-star participants, those with 3 stars had lower risk-adjusted mortality (4.3% versus 11.1%) and risk-adjusted morbidity (47.6% versus 19.8%) during July 2012 – June 2015.



2b1.4. What is your interpretation of the results in terms of demonstrating validity? (i.*e., what do the results mean and what are the norms for the test conducted?*)
The test results show that the composite measure behaves as expected and that results are reasonably consistent across two consecutive overlapping time periods. These results support the validity of the composite measure as a quality measure for MVRR + CABG procedures.

2b2. EXCLUSIONS ANALYSIS

Note: *Applies to the composite performance measure, as well all component measures unless they are already endorsed or are being submitted for individual endorsement.*

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*)
N/A

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*)
N/A

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*)
N/A

2b3. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES
Note: *Applies to all outcome or resource use component measures, unless already endorsed or are being submitted for individual endorsement.*
If not an intermediate or health outcome, or PRO-PM, or resource use measure, skip to section 2b4.

2b3.1. What method of controlling for differences in case mix is used? *(check all that apply)*

☒ **Endorsed (or submitted) as individual performance measures**

☐ **No risk adjustment or stratification**

☒ **Statistical risk model with 55 risk factors**

☐ **Stratification by risk categories**

☐ **Other,**

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

Please see 2b3.3a and 2b3.4a below.

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

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

To adjust for case mix in the STS MVRR + CABG Composite Score [1], the published 2008 STS valve +CABG model [2] was modified and re-estimated in the current study population. The main reason for modifying the model was to be able to calculate predicted risk estimates for patients in the current study population who did not meet inclusion/exclusion criteria for the existing 2008 STS valve + CABG model. In addition, although the existing STS models predict the endpoints of "operative mortality" and "operative mortality or major morbidity" there is no existing model for predicting "major morbidity" as defined in the current study. In the future, as STS risk models are revised over time, the composite measure will be calculated with the most up to date STS risk model for the MVRR + CABG population.

Except where noted below, covariates for the modified operative mortality model were identical to the STS 2008 operative mortality model and covariates for the new major morbidity model were identical to the STS 2008 operative mortality or major morbidity model. For each of the two models, the list of covariates was modified as follows.

- Adjust for concomitant tricuspid repair. The STS 2008 models excluded patients undergoing a concomitant tricuspid procedure. Because the current study included patients undergoing concomitant tricuspid repair, an indicator variable for tricuspid repair was included.
- Adjust for tricuspid insufficiency using categories none or mild, moderate, and severe. The 2008 models included indicators of at least moderate tricuspid insufficiency. Because of the inclusion of operations with concurrent TV repair procedures, the surgeon panel felt it was necessary to more

finely adjust the degrees of tricuspid insufficiency. The modified models include separate indicator variables for moderate tricuspid insufficiency and severe tricuspid insufficiency.

- Adjust for infectious endocarditis using categories active, treated, and none. The 2008 models include an indicator for active infections endocarditis but not for treated infectious endocarditis. The modified models include separate indicator variables for treated infectious endocarditis and active infectious endocarditis.

Considerations for adjusting for tricuspid repair

It is a generally accepted principle not to use what may be discretionary procedural decisions (e.g., whether or not to add a tricuspid valve repair) in profiling models. However, as discussed in the main article, there is accumulating evidence of the potential longitudinal merits of concomitant TVr. and the surgeon panel wanted to avoid discouraging the performance of this procedure by failing to account for its increased inherent risk of morbidity. Furthermore, the panel felt that the need to perform TVr may be a proxy for more advanced disease that may not be captured perfectly in the current STS data collection form.

References

1. Rankin JS, Badhwar V, He X, Jacobs JP, Gammie JS, Furnary AP, Fazzalari FL, Han J, O'Brien SM, Shahian DM. The Society of Thoracic Surgeons Mitral Valve Repair/Replacement plus Coronary Artery Bypass Grafting Composite Score: A Report of The Society of Thoracic Surgeons Quality Measurement Task Force – This manuscript is currently being prepared for submission to The Annals of Thoracic Surgery.
2. Shahian DM, O'Brien SM, Filardo G, Ferraris VA, et al. The Society of Thoracic Surgeons 2008 cardiac surgery risk models: part 3--valve plus coronary artery bypass grafting surgery. Ann Thorac Surg 2009 Jul;88(1 Suppl):S43-62.

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)**

Expert group consensus

2b3.4a. What were the statistical results of the analyses used to select risk factors? Estimated odds ratios from the modified STS 2008 models are summarized in the table below.

Effect	Morbidity: OR (95% CI)	Morbidity: P-value	Mortality: OR (95% CI)	Mortality: P-value
Effects that do not interact with MV repair/replacements	*	*	*	*
Preoperative atrial fibrillation	1.09 (1.02, 1.17)	0.0125	1.04 (0.92, 1.18)	0.4926
Race (v. others)	*	*	*	*
Black	1.21 (1.08, 1.35)	0.0007	*	*
Hispanic	1.16 (1.00, 1.35)	0.0529	*	*
CVD (v. no)	*	*	*	*

Effect	Morbidity: OR (95% CI)	Morbidity: P-value	Mortality: OR (95% CI)	Mortality: P-value
CVD with CVA	1.21 (1.10, 1.33)	0.0001	1.01 (0.86, 1.19)	0.9223
CVD without CVA	1.09 (0.98, 1.21)	0.1264	*	*
Number Diseased Vessels (3 v. 2, 2 v. 1/0)	1.16 (1.11, 1.21)	<.0001	1.16 (1.08, 1.26)	<.0001
Pre-op IABP or inotrope	2.21 (1.98, 2.47)	<.0001	1.43 (1.22, 1.69)	<.0001
Hypertension	1.11 (1.02, 1.20)	0.0189	*	*
Immunosuppressive treatment	1.17 (1.02, 1.34)	0.0264	1.29 (1.02, 1.63)	0.0303
Peripheral vascular disease	1.08 (1.00, 1.17)	0.0536	1.28 (1.11, 1.48)	0.0007
MI (v. no recent MI)	*	*	*	*
1-21 days	1.32 (1.23, 1.42)	<.0001	1.30 (1.13, 1.50)	0.0002
<=24 hrs	1.48 (1.16, 1.89)	0.0015	1.76 (1.28, 2.40)	0.0004
Number of previous operations (v. 0)	*	*	*	*
1 previous operation	1.45 (1.15, 1.83)	0.0017	2.79 (1.88, 4.14)	<.0001
2 or more previous operations	1.50 (1.00, 2.24)	0.0485	2.68 (1.41, 5.06)	0.0025
Diabetes (v. no)	*	*	*	*
Non-insulin diabetes	1.22 (1.12, 1.32)	<.0001	1.35 (1.17, 1.57)	<.0001
Insulin diabetes	1.08 (1.01, 1.16)	0.0233	1.10 (0.97, 1.24)	0.1565
Chronic lung disease (severe v moderate, or moderate v none-mild)	1.10 (1.07, 1.14)	<.0001	1.16 (1.10, 1.22)	<.0001
Dialysis v. no dialysis & creatinine = 1.0	2.17 (1.88, 2.50)	<.0001	2.66 (2.19, 3.23)	<.0001
Creatinine per 1 unit increase	1.62 (1.51, 1.73)	<.0001	1.46 (1.33, 1.61)	<.0001
Female (at BSA=1.8) v. male (at BSA=2.0)	1.20 (1.11, 1.29)	<.0001	1.39 (1.21, 1.59)	<.0001
Status (v. elective)	*	*	*	*
Urgent	1.26 (1.18, 1.36)	<.0001	1.09 (0.96, 1.24)	0.1821
Emergent - no resuscitation	2.53 (1.75, 3.65)	<.0001	1.74 (1.12, 2.73)	0.0148
Emergent+resuscitation/Emergent Salvage	1.90 (1.07, 3.38)	0.0292	5.13 (2.83, 9.31)	<.0001
Active infections endocarditis	1.48 (1.20, 1.83)	0.0003	1.63 (1.18, 2.24)	0.0027
Treated infections endocarditis	0.91 (0.72, 1.16)	0.4538	0.57 (0.33, 0.97)	0.0393

Effect	Morbidity: OR (95% CI)	Morbidity: P-value	Mortality: OR (95% CI)	Mortality: P-value
Body surface area, m^2	*	*	*	*
1.6 v. 2.0 in male	1.16 (1.01, 1.34)	0.0354	1.32 (1.02, 1.72)	0.0354
1.8 v. 2.0 in male	1.02 (0.97, 1.08)	0.4400	1.07 (0.97, 1.17)	0.1703
2.2 v. 2.0 in male	1.09 (1.05, 1.14)	<.0001	1.08 (1.01, 1.16)	0.0234
1.6 v. 1.8 in female	1.12 (1.06, 1.18)	0.0002	1.24 (1.12, 1.36)	<.0001
2.0 v. 1.8 in female	1.06 (1.00, 1.12)	0.0360	1.03 (0.94, 1.12)	0.5595
2.2 v. 1.8 in female	1.33 (1.15, 1.54)	0.0002	1.34 (1.06, 1.68)	0.0133
Time trend (half year increase)	0.98 (0.96, 1.00)	0.0541	1.03 (1.00, 1.06)	0.0440
Left main disease	*	*	1.09 (0.96, 1.24)	0.1778
Unstable angina (no MI < 8days)	*	*	1.01 (0.87, 1.17)	0.9382
Mitral stenosis	*	*	1.21 (1.01, 1.46)	0.0399
Mitral insufficiency (>= moderate)	0.95 (0.86, 1.05)	0.3396	*	*
Moderate tricuspid insufficiency (v. no-mild)	1.10 (1.02, 1.20)	0.0189	1.10 (0.96, 1.26)	0.1618
Severe tricuspid insufficiency (v. no-mild)	1.12 (0.98, 1.29)	0.1051	1.12 (0.89, 1.41)	0.3448
Mitral valve repair (v. replacement)	0.69 (0.59, 0.81)	<.0001	0.81 (0.59, 1.10)	0.1784
Tricuspid valve repair (v. none)	1.33 (1.19, 1.49)	<.0001	1.04 (0.85, 1.27)	0.7010
Effects that interacts with procedure groups and were modeled separately for MV replacement and MV repairs In MV replacements + CABG	*	*	*	*
Age	*	*	*	*
60 v. 50 (no reoperations, non-emergent)	1.16 (1.09, 1.23)	<.0001	1.70 (1.51, 1.91)	<.0001
70 v. 50 (no reoperations, non-emergent)	1.35 (1.20, 1.52)	<.0001	2.88 (2.28, 3.64)	<.0001
80 v. 50 (no reoperations, non-emergent)	1.57 (1.34, 1.84)	<.0001	4.84 (3.62, 6.49)	<.0001
Congestive heart failure (v. no)	*	*	*	*
CHF not NYHA IV	1.15 (1.04, 1.28)	0.0063	1.14 (0.94, 1.37)	0.1794
CHF NYHA IV	1.36 (1.18, 1.55)	<.0001	1.49 (1.21, 1.83)	0.0002
Ejection fraction per 10-unit decrease	1.12 (1.09, 1.16)	<.0001	1.04 (0.96, 1.14)	0.3436

Effect	Morbidity: OR (95% CI)	Morbidity: P-value	Mortality: OR (95% CI)	Mortality: P-value
Shock	2.07 (1.59, 2.69)	<.0001	1.89 (1.49, 2.39)	<.0001
In MV repairs + CABG	*	*	*	*
Age	*	*	*	*
60 v. 50 (no reoperations, non-emergent)	1.16 (1.10, 1.21)	<.0001	1.45 (1.31, 1.61)	<.0001
70 v. 50 (no reoperations, non-emergent)	1.34 (1.21, 1.47)	<.0001	2.11 (1.72, 2.60)	<.0001
80 v. 50 (no reoperations, non-emergent)	1.55 (1.36, 1.76)	<.0001	3.04 (2.35, 3.92)	<.0001
Congestive heart failure (v. no)	*	*	*	*
CHF not NYHA IV	1.15 (1.05, 1.27)	0.0027	1.27 (1.06, 1.51)	0.0087
CHF NYHA IV	1.32 (1.18, 1.49)	<.0001	1.40 (1.14, 1.73)	0.0016
Shock	1.97 (1.56, 2.47)	<.0001	1.89 (1.49, 2.39)	<.0001
Ejection fraction per 10-unit decrease	1.12 (1.09, 1.16)	<.0001	1.13 (1.06, 1.21)	0.0002

CHF = congestive heart failure; CVA = cerebrovascular accident (stroke); CVD = cardiovascular disease; EF = ejection fraction; IABP = intra-aortic balloon pump; MI = myocardial infarction; NA = variable not used in model and estimate not available; NYHA = New York Heart Association.

*cell intentionally left blank

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.

Please see our response in 1.8 above, including explanation for the continued inclusion of race in the STS Adult Cardiac risk models.

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*)

The modified models were assessed using data from 26,355 patients undergoing MVRR + CABG during July 2011 – June 2014.

Discrimination

To gauge discrimination, we calculated the c-statistics of both models. Bootstrapping was used to estimate and adjust for the "optimism" from estimating and evaluating the model on the same sample [1].

Calibration

The model fit was evaluated using 5-fold cross validation. The entire sample was randomly split into five equal sized groups. The calibration plot was created by following these steps:

1. One of the five groups was used as the testing sample
2. The other four groups were combined into the training sample
3. The revised model was estimated using the training sample
4. The expected probability of experience the event in the testing sample was calculated using the model estimated in step 3.
5. The expected probability (from step 4) and observed event rates were then compared in the testing sample and the calibration plot was created.

The above five steps were repeated five times so that each group was used as the testing sample once. In the end, we had five calibration plots for each model.

Reference

1. Harrell, F. E., Kerry L. Lee, and Daniel B. Mark. Tutorial in biostatistics multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Statistics in medicine 15 (1996): 361-387.

Provide the statistical results from testing the approach to controlling for differences in patient characteristics (case mix) below.
If stratified, skip to 2b3.9

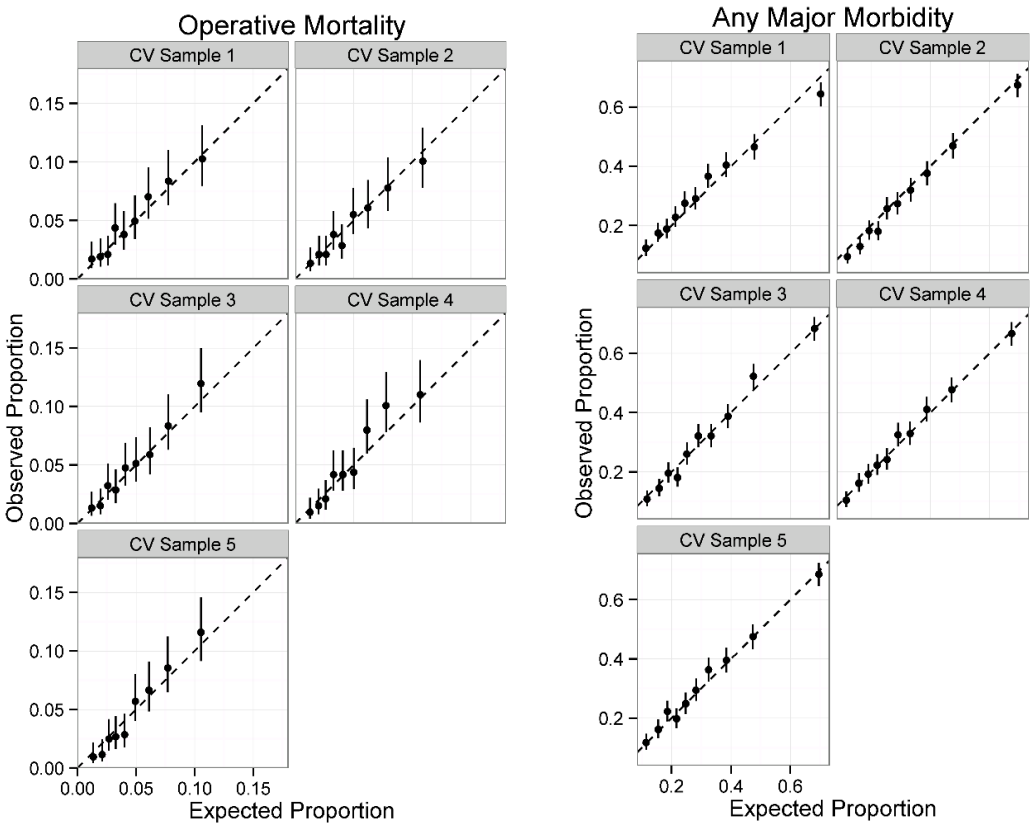
2b3.6. Statistical Risk Model Discrimination Statistics (*e.g., c-statistic, R-squared*)**:**

The bootstrap-adjusted C statistic was 0.708 for the morbidity model and 0.738 for the mortality model. These numbers were comparable to the STS 2008 models when evaluated using the same sample (0.707 and 0.738 for morbidity and mortality endpoints, respectively.)

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

N/A. The Hosmer-Lemeshow statistic was not calculated.

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



Plots of observed versus expected in cross validation samples, operative mortality

2b3.9. Results of Risk Stratification Analysis:

N/A

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

The results demonstrated that the STS cardiac surgery risk models are well calibrated and have good discrimination power. They are suitable for controlling differences in case-mix between centers.

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

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

Note: *Applies to the composite performance measure.*

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 degree of uncertainty surrounding an STS participant's composite measure estimate is indicated by calculating 95% Bayesian credible intervals (CI's) which are similar to conventional confidence intervals. Point estimates and CI's for an individual STS participant are reported along with a comparison to various benchmarks based on the national sample. Benchmarks include the overall average STS composite score and several percentiles (minimum, 10th, 25th, 75th, 90th, maximum). In addition, the composite measure result is converted into categories labeled as 1 to 3 stars. An STS participant receives 2 stars if the Bayesian credible interval surrounding their composite score overlaps the overall STS average. This rating implies that the STS participant's performance was not statistically different from the overall STS national average. If the Bayesian CI falls entirely above the STS national average, the participant receives 3 stars (higher-than-expected performance). If the Bayesian CI falls entirely below the STS national average, the participant receives 1 star (lower-than-expected performance).

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*)
Among participants with at least 25 cases over 3 years, around 91% of participants have received 2 stars, and the remaining participants have received either 1 or 3 stars.

Performance categories

July 2011 – June 2014

Category	All Participants	Participants N≥ 25
*	**Number of Participants, %**	**Number of Participants, %**
1-star	14, 2.0%	8, 2.3%
2-star	666, 94.7%	310, 90.9%
3-star	23, 3.3%	23, 6.7%

*cell intentionally left blank

July 2012 – June 2015

Category	All Participants	Participants N≥ 25
*	**Number of Participants, %**	**Number of Participants, %**
1-star	10, 1.5%	10, 2.9%
2-star	657, 95.5%	314, 91.0%
3-star	21, 3.1%	21, 6.1%

*cell intentionally left blank

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

The Bayesian methodology allows direct probability interpretation of the results. The identified differences in performance are both statistically significant and clinically meaningful. The surgeon panel and users are satisfied with the distribution of participants across performance categories.

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

Note: *Applies to all component measures, unless already endorsed or are being submitted for individual endorsement.*

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

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

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

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

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

2b6. MISSING DATA ANALYSIS AND MINIMIZING BIAS

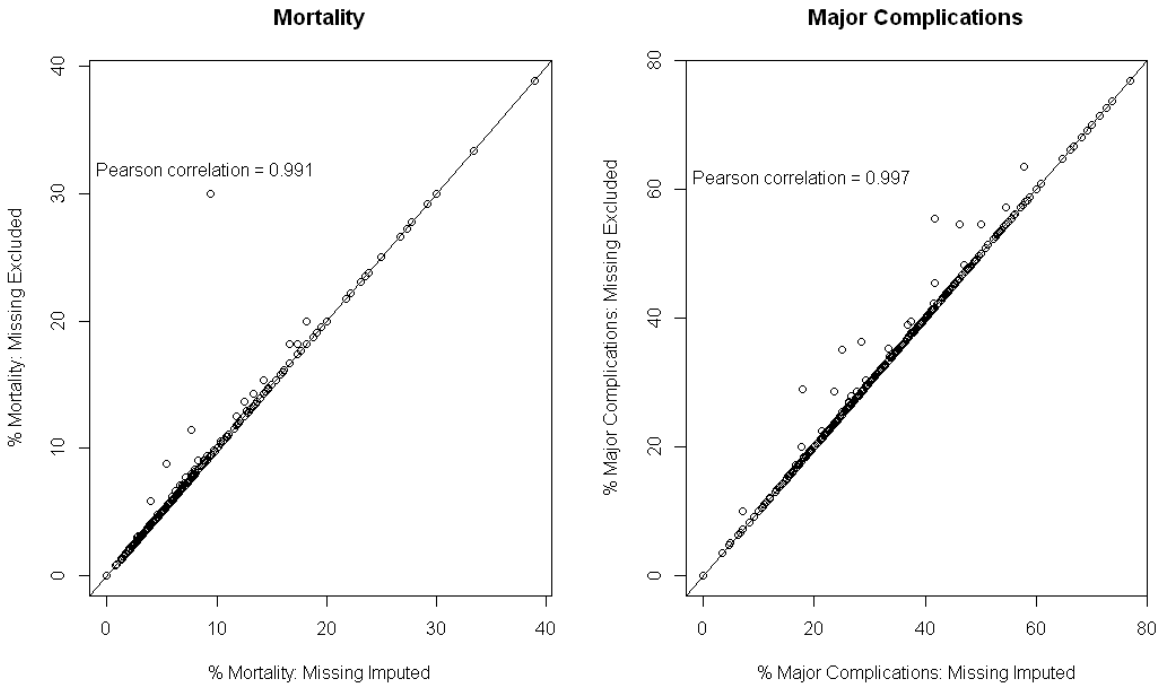
Note: *Applies to the overall composite measure.*

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*)
Missing data for risk model covariates was extremely rare: All model predictors had <5% missing and the majority had <1% missing. Missing data occurred in 0.55% of records for operative mortality and 0.44% of records for major complications. In the rare case of missing data, unknown values were imputed to the most common category of categorical variables and to the median or subgroup-specific median of continuous variables. Single imputation was used in the multivariable models consistent with the STS methodology used in the creation of the original STS model. More computationally intensive missing data strategies, such as multiple imputation, were not used for this analysis because of the low rate of missing data and because it would be impractical to implement them in combination with the computationally intensive Bayesian Markov Chain Monte Carlo (MCMC) procedure used for estimation of composite scores. Moreover, the use of multiple imputation has had negligible impact in previous STS analyses with similar low rates of missing data. For a

comparison of single versus multiple imputation results in the development of the STS 2008 risk model, please see http://people.duke.edu/~obrie027/STS2008/. In that analysis, using multiple imputation did not appreciably widen the confidence intervals around model estimates. Moreover, any differences in point estimates were small relative to their standard error. Similar results have been found in a number of STS publication analyses.

A 30-day vital status category of "unknown" is available for those instances (e.g., homeless patients) in which the status of the patient cannot be ascertained despite good faith efforts to do so. In order to prevent excessive or inappropriate use of this vital status category, stringent new limitations on the use of this category were implemented in 2016, retroactive to 2015 data. This will further assure the accuracy of the operative mortality endpoint, which includes a small percentage of patients who die between hospital discharge and 30 days.

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*)
The overall frequency of missing data was 0.55% for operative mortality and 0.44% for major complications. The median participant-specific frequency of missing data was 0% (range 0% to 65%) for mortality and 0% (range 0% to 40%) for major complications. The percent of participants with >10% missing data was 0.7% for mortality and 1.5% for major complications. As a sensitivity analysis, we re-calculated participant-specific mortality and complication rates after excluding records with missing data from the denominator. As shown in the figure below, there was high (>0.99) correlation between participant-specific rates calculated with missing data excluded versus imputed.



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*)

These results suggest that our handling of missing outcome data is unlikely to impact performance results for the vast majority of participants.

2c. EMPIRICAL ANALYSIS TO SUPPORT COMPOSITE CONSTRUCTION APPROACH

Note: *If empirical analyses do not provide adequate results—or are not conducted—justification must be provided and accepted in order to meet the must-pass criterion of Scientific Acceptability of Measure Properties. Each of the following questions has instructions if there is no empirical analysis.*

2d1. Empirical analysis demonstrating that the component measures fit the quality construct, add value to the overall composite, and achieve the object of parsimony to the extent possible.

2d1.1 Describe the method used (*describe the steps—do not just name a method; what statistical analysis was used; if no empirical analysis, provide justification)*
To verify that each domain contributes statistical information but does not dominate the composite, we calculated the correlations between each domain-specific estimate and the overall composite score. These analyses were performed using data from July 2011 – June 2014.

2d1.2. What were the statistical results obtained from the analysis of the components? (e.g., *correlations, contribution of each component to the composite score, etc*.; *if no empirical analysis, identify the components that were considered and the pros and cons of each*)

Pearson Correlation With Overall Composite: Mortality	Pearson Correlation With Overall Composite: Morbidity
0.60	0.91

The Pearson correlations were 0.60 for mortality versus the overall composite measure and 0.91 for morbidity domain score versus overall score.

2d1.3. What is your interpretation of the results in terms of demonstrating that the components included in the composite are consistent with the described quality construct and add value to the overall composite? (i*.e., what do the results mean in terms of supporting inclusion of the components; if no empirical analysis, provide rationale for the components that were selected)*

Although risk-adjusted morbidity explains more of the variation in the overall composite score, it does not dominate. Both domains also contribute statistical information.

2d2. Empirical analysis demonstrating that the aggregations and weighting rules are consistent with the quality construct and achieve the objective of simplicity to the extent possible

2d2.1 Describe the method used (*describe the steps—do not just name a method; what statistical analysis was used; if no empirical analysis, provide justification)*

To form the composite, we rescaled the morbidity and mortality domains by dividing by their respective standard deviations across STS participants and then added the two domains together. This weighting was then assessed by an expert panel to determine if it provided an appropriate reflection of the relative importance of the two domains. To facilitate the assessment, we calculated for a 1 percentage point change in

mortality, what percentage point change in morbidity would be needed to achieve the same impact on the composite measure.

2d2.2. What were the statistical results obtained from the analysis of the aggregation and weighting rules? (e.g., *results of sensitivity analysis of effect of different aggregations and/or weighting rules; if no empirical analysis, identify the aggregation and weighting rules that were considered and the pros and cons of each*)

After rescaling, the relative weights in the final composite of risk-standardized mortality and risk-standardized major morbidity were 0.74 and 0.26, respectively. An implication of this weighting is that a 1 percentage point change in a participant's risk-adjusted mortality rate has the same impact as a 2.8 percentage point change in the site's risk-adjusted morbidity rate.

2d2.3. What is your interpretation of the results in terms of demonstrating the aggregation and weighting rules are consistent with the described quality construct? (i.*e., what do the results mean in terms of supporting the selected rules for aggregation and weighting; if no empirical analysis, provide rationale for the selected rules for aggregation and weighting*)

This weighting was consistent with our expert panel's clinical assessment of each domain's relative importance.

3. Feasibility

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

3a. Byproduct of Care Processes

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

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

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

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**.

Some data elements are in defined fields in 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).

The STS Adult Cardiac Surgery Database (ACSD) has more than 1,030 participants as of August 2020, and local availability of data elements in electronic format will vary across institutions. Some institutions may have full EHR capability while others may have partial, or no availability. However, all data elements from participating institutions are submitted to the STS ACSD in electronic format following a standard set of data specifications. The majority of participating institutions obtain data entry software products that are certified for the purposes of collecting STS ACSD data elements. The STS Adult Cardiac Surgery Database (ACSD) has more than 1,100 participants, and local availability of data elements in electronic format will vary across institutions. Some institutions may have full EHR capability while others may have partial, or no availability. However, all data elements from participating institutions are submitted to the STS ACSD in electronic format following a standard set of data specifications. The majority of participating institutions obtain data entry software products that are certified for the purposes of collecting STS ACSD data elements.

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.

The data elements included in this measure have been standard in the STS Adult Cardiac Surgery Database for at least 6 years and some of them have been part of the database for more than 20 years. The variables are considered to be data elements that are readily available and already collected as part of the process of providing care.

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)*.

Data Collection:

There are no additional costs for data collection specific to this measure for those presently using and participating in the STS Adult Cardiac Surgery Database. Costs to develop and maintain the measure included volunteer cardiothoracic surgeon time, STS staff time, and Duke Clinical Research Institute statistician and project management time.

Other fees:

STS Adult Cardiac Surgery Database participants (generally a group of surgeons) pay annual participant fees of $3,500 or $4,750, depending on whether the majority of surgeons in a participant group are STS members. As a benefit of STS membership, the member-majority participants are charged the lesser of the two fees. Also, member-majority participants pay an additional fee of $150 per surgeon; non-member-majority participants pay an additional fee of $350 per surgeon. Data Collection:

There are no additional costs for data collection specific to this measure for those presently using and participating in the STS Adult Cardiac Surgery Database. Costs to develop and maintain the measure included volunteer cardiothoracic surgeon time, STS staff time, and Duke Clinical Research Institute statistician and project management time.

Other fees:

STS Adult Cardiac Surgery Database participants (single cardiothoracic surgeons or a group of surgeons) pay annual participant fees of $3,500 or $4,750, depending on whether the majority of surgeons in a participant group are STS members. As a benefit of STS membership, the member-majority participants are charged the lesser of the two fees. Also, member-majority participants pay an additional fee of $150 per surgeon; non-member-majority participants pay an additional fee of $350 per surgeon.

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 STS Public Reporting https://www.sts.org/registries/sts-public-reporting Quality Improvement (Internal to the specific organization) STS Adult Cardiac Surgery Database https://www.sts.org/registries-research-center/sts-national-database/adult-cardiac-surgery-database

*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

Voluntary STS Public Reporting – approximately 79% of STS Adult Cardiac Surgery Database participants are enrolled as of October 2020.
This measure has been publicly reported since 2017.
(https://publicreporting.sts.org/acsd)
STS Adult Cardiac Surgery Database Participant Feedback Reports provide performance results for this measure to participants. (see details in 4a2.1.1 below)

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

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

N/A

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.

As of November 2020, there are 1,030 active U.S. and Canadian participants in the STS Adult Cardiac Surgery Database (ACSD). A "participant" is generally a group of cardiothoracic surgeons who agree to submit case records for analysis and comparison with benchmarking data for quality improvement initiatives. At the option of the surgical group, the ACSD participant can include a hospital and/or associated anesthesiologists. It is for this reason that we have indicated (on the Specifications tab, question #S.20) that this measure is specified/tested for both the "clinician: group/practice" and "facility" levels of analysis.

(For more information on STS "participants," see our response to 1.5 in the measure testing form.)

All ACSD participants receive quarterly data reports with their performance results, reported in an easy-to-understand format. The participant's score is illustrated graphically in relation to the 25th, 50th and 75th percentiles of the distribution across all participants who were eligible for inclusion in that quarter´s analysis, and is also accompanied by the 95% Bayesian credible interval. Surgeons easily grasp this result and the visual display clearly illustrates how they perform compared to their peers on a quarterly basis. In addition, these risk-adjusted results allow surgeons to compare their patients´ outcomes with national benchmarks and to initiate quality improvement efforts as needed.

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.

Please see response under 4a2.1.1

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.

The adult cardiac surgeons from across the U.S. who comprise the STS Adult Cardiac Surgery Task Force meet periodically to discuss the participant reports and to consider potential enhancements to the ACSD. Additions/clarifications to the data collection form and to the content/format of the participant reports are discussed and implemented as appropriate.

Most recently, STS surgeon members have expressed interest in real-time, online data updates, which has led to the development of dashboard-type reporting on STS.org. Developed by IQVIA, the Society´s new data warehouse (https://www.sts.org/registries-research-center/sts-national-database/database-transition-resources), the new platform for the Adult Cardiac Surgery Database was released in early 2020. Surgeon members have access to near-real time data updates in the dashboard. Enhancements to dashboard functionality are ongoing.

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

Please see response under 4a2.2.1

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

Voluntary participation in ACSD public reporting has continually increased over the years that the initiative has been available, from 38% of ACSD participants in 2014, to 49% in 2016, to 67% in 2018, to approximately 79% in October 2020. This trend suggests that feedback from ACSD participants and others who access the performance data available on STS.org is sufficiently positive to promote ever-increasing participation in public reporting.

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

N/A

Improvement

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

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

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

Please see table below displaying 2017-2019 star ratings for this measure, in percentages.

The data demonstrate that the trend for the MVRR+CABG measure since 2017 (the year the MVRR and MVRR+CABG composites were introduced) is not consistent with the general trend seen for other STS composite measures – a decrease over time in the percentage of surgical programs with 1-star and 3-star ratings and a corresponding increase in 2-star programs. With additional years of experience with this composite, we anticipate decreased variation in performance and a higher percentage of participants in the STS Adult Cardiac Surgery Database to be rated in the 2-star (or "as expected") category.

	Stars	2019	2018	2017
MVRR + CABG	*	2.55	2.08	2.74
	** 88.0	89.97	91.78	
	***	9.45	7.96	5.48

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.

All public reporting initiatives have the potential for unintended consequences, including gaming and risk aversion. We attempt to control the former through a careful audit process; 10% of STS Adult Cardiac Surgery Database participants were audited in each year from 2014 through 2019. (Our audit plans for 2020 were canceled due to the coronavirus pandemic; we expect to resume with 10% audits in 2021.) We control for risk aversion by having a robust methodology that appropriately adjusts the expected risk for providers who care for sicker patients.

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

N/A

5. Comparison to Related or Competing Measures

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

5. Relation to Other NQF-endorsed Measures

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

Yes

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

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

Related measures (not listed in drop-down menu for 5.1a):

0696 - STS CABG Composite

2561 - Aortic Valve Replacement Composite Score

2563 - Aortic Valve Replacement + CABG Composite Score

3031 - Mitral Valve Repair/Replacement Composite Score

5a. Harmonization of Related Measures

The measure specifications are harmonized with related measures;
OR
The differences in specifications are justified

5a.1. If this measure conceptually addresses EITHER the same measure focus OR the same target population as NQF-endorsed measure(s):
Are the measure specifications harmonized to the extent possible?

Yes

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

N/A

5b. Competing Measures

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

5b.1. If this measure conceptually addresses both the same measure focus and the same target population as NQF-endorsed measure(s):
Describe why this measure is superior to competing measures (e.g., a more valid or efficient way to measure quality); OR provide a rationale for the additive value of endorsing an additional measure. (Provide analyses when possible.)

N/A

Appendix

A.1 Supplemental materials may be provided in an appendix. All supplemental materials (such as data collection instrument or methodology reports) should be organized in one file with a table of contents or bookmarks. If material pertains to a specific submission form number, that should be indicated. Requested

information should be provided in the submission form and required attachments. There is no guarantee that supplemental materials will be reviewed.

Attachment Attachment: STS_MVRR_-_CABG_Composite_Score_Appendix_-_S.4-11-14-15_1b.2-_1b.4-_102020-637408793378385811.pdf

Contact Information

Co.1 Measure Steward (Intellectual Property Owner): The Society of Thoracic Surgeons

Co.2 Point of Contact: Mark, Antman, mantman@sts.org, 312-202-5856-

Co.3 Measure Developer if different from Measure Steward: The Society of Thoracic Surgeons

Co.4 Point of Contact: Mark, Antman, mantman@sts.org, 312-202-5856-

Additional Information

Ad.1 Workgroup/Expert Panel involved in measure development

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

The STS Quality Measurement Task Force (chaired by David Shahian, MD) is responsible for measure development. Members of the STS Task Force on Quality Initiatives provide clinical expertise as needed. The STS Workforce on Quality meets at the STS Annual Meeting and reviews the measures on a yearly basis. Changes or updates to the measure will be at the recommendation of the Workforce.

Quality Measurement Task Force

David M. Shahian, MD, Chair; Massachusetts General Hospital & Harvard Medical School, Boston, MA

Diane Alejo; Johns Hopkins Univ., Baltimore, MD

Vinay Badhwar, MD; West Virginia University Hospitals, Morgantown, WV

Jordan Bloom, MD; Massachusetts General Hospital, Boston, MA

Michael Bowdish, MD; Torrance Memorial Medical Center, Los Angeles, CA

Joseph Cleveland, Jr., MD; University of Colorado Anschutz Medical Campus, Aurora, Co

Nimesh Desai, MD; Hospital of the University of Pennsylvania, Philadelphia, PA

James Edgerton, MD; Cardiac Surgery Specialists, Plano, TX

Fred Edwards, MD; University of Florida College of Medicine, Jacksonville, FL

Melanie Edwards, MD; Saint Joseph Mercy Health System, Ypsilanti, MI

Vic Ferraris, MD; University of Kentucky Medical Center, Lexington, KY

Anthony Furnary, MD; Providence Alaska Medical Center, Anchorage, AK

Joshua Goldberg, MD; Westchester Medical Center, Valhalla, NY

Jeffrey P. Jacobs, MD; University of Florida, Gainesville, FL

Marshall Jacobs, MD; Johns Hopkins Cardiac Surgery, Baltimore, MD

Karen Kim, MD; Univ. of Michigan Hospitals & Health Centers, Ann Arbor, MI

Benjamin Kozower, MD; Washington University School of Medicine, St. Louis, MO

Paul Kurlansky, MD; Columbia HeartSource/Columbia University Medical Center, New York, NY

Kevin Lobdell, MD; Atrium Health, Charlotte, NC

Mitchell Magee, MD; Southwest Cardiothoracic Surgeons, Dallas, TX

Gaetano Paone, MD; Henry Ford Hospital, Detroit, MI

J. Scott Rankin, MD; WVU Heart & Vascular Institute, West Virginia University, Morgantown, WV

Charles Schwartz, MD; St. Joseph Mercy Hospital, Pontiac, MI

Vinod Thourani, MD; MedStar Washington Hospital Center, Washington, DC

Christina Vassileva, MD; U Mass Memorial Medical Center, Worcester, MA

Moritz Wyler von Ballmoos, MD; Houston Methodist DeBakey Heart & Vascular Center, Houston, TX

Sean M. O'Brien, PhD; Duke Clinical Research Institute, Durham, NC

Measure Developer/Steward Updates and Ongoing Maintenance

Ad.2 Year the measure was first released: 2016

Ad.3 Month and Year of most recent revision: 06, 2016

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

Ad.5 When is the next scheduled review/update for this measure? 01, 2021

Ad.6 Copyright statement: N/A

Ad.7 Disclaimers: N/A

Ad.8 Additional Information/Comments: N/A