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 #: 3030

Corresponding Measures:

De.2. **Measure Title:** STS Individual Surgeon Composite Measure for Adult Cardiac Surgery

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

De.3. **Brief Description of Measure:** The STS Individual Surgeon Composite Measure for Adult Cardiac Surgery includes five major procedures (isolated CABG, isolated AVR, AVR+CABG, MVRR, MVRR+CABG) and comprises the following two domains:

Domain 1 – Risk-Adjusted Operative Mortality

Operative mortality is defined as death before hospital discharge or within 30 days of the operation.

Domain 2 – Risk-Adjusted 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.

All measures are based on audited clinical data collected in the STS Adult Cardiac Surgery Database. Individual surgeons with at least 100 eligible cases during the 3-year measurement window will receive a score for each domain and an overall composite score. In addition to calculating composite score point estimates with credible intervals, surgeons will be assigned 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 in detail this multiprocedural, multidimensional composite measure.

The STS Individual Surgeon Composite Measure for Adult Cardiac Surgery includes five major procedures, i.e., isolated coronary artery bypass grafting (CABG), isolated aortic valve replacement (AVR), AVR+CABG, isolated mitral valve repair or replacement (MVRR), and MVRR+CABG, and comprises the following two domains:

Domain 1 – Risk-Adjusted Operative Mortality

Operative mortality is defined as death before hospital discharge or within 30 days of the operation.

Domain 2 – Risk-Adjusted 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

Individual surgeons with at least 100 eligible cases during the 3-year measurement window will receive a score for each domain and an overall composite score. In addition to calculating composite score point estimates with credible intervals, surgeons will be assigned 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 undergo isolated CABG, isolated AVR, AVR+CABG, isolated MVRR, and MVRR+CABG.

Time Window: 3 years

By including composite performance scores for a portfolio of five procedures that account for nearly 80% of a typical STS Adult Cardiac Surgery Database participant surgeon's clinical activity, this metric provides a more balanced and comprehensive perspective than focusing on just one procedure or one end point. Recognizing that surgeons' practices vary, each surgeon's composite performance is implicitly "weighted" by the proportion of each type of procedure he or she performs. For instance, the results of surgeons who primarily perform mitral procedures are affected most by their mitral surgery results. This approach is especially relevant for surgeons with highly specialized practices who may do relatively few isolated CABG procedures and whose performance would thus be difficult to assess using a CABG measure only. Finally, performance on each of these procedures is estimated using risk models specific to those procedures, in most cases the exact or slightly modified versions of previously published models (references provided below).

Final Composite Score:

The overall composite score was calculated as a weighted sum of (1 minus risk-adjusted mortality rate) and (1 minus risk-adjusted major morbidity rate). Mortality and morbidity rates were weighted inversely by their respective standard deviations across surgeons. This procedure is equivalent to first rescaling mortality and morbidity rates by their respective standard deviations across surgeons and then assigning equal weighting to the rescaled mortality rate and rescaled morbidity rate. Standard deviations derived from the data were used to define the final composite measure as 0.81 x (1 minus risk-standardized mortality rate) + 0.19 x (1 minus risk-standardized complication rate).

Details regarding the current STS adult cardiac surgery risk models can be found in the following manuscripts:

- Shahian DM, O´Brien SM, Filardo G, Ferraris VA, et al. The Society of Thoracic Surgeons 2008 cardiac surgery risk models: part 1--coronary artery bypass grafting surgery. Ann Thorac Surg. 2009 Jul;88(1 Suppl):S2-22.

- O'Brien SM, Shahian DM, Filardo G, et al. The Society of Thoracic Surgeons 2008 cardiac surgery risk models: part 2—isolated valve surgery. Ann Thorac Surg 2009;88(1 Suppl):S23–42.
- 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.

Additional details regarding the Individual Surgeon Composite Measure for Adult Cardiac Surgery are provided in the attached manuscript:

Shahian DM, He X, Jacobs JP, Kurlansky PA, Badhwar V, Cleveland JC Jr, Fazzalari FL, Filardo G, Normand SL, Furnary AP, Magee MJ, Rankin JS, Welke KF, Han J, O´Brien SM. The Society of Thoracic Surgeons Composite Measure of Individual Surgeon Performance for Adult Cardiac Surgery: A Report of The Society of Thoracic Surgeons Quality Measurement Task Force. Ann Thorac Surg. 2015;100:1315-25.

S.6. Denominator Statement: See response in S.4. Numerator Statement

Patient Population: The analysis population consists of patients aged 18 years or older who undergo isolated CABG, isolated AVR, AVR+CABG, isolated MVRR, and MVRR+CABG.See response in S.4. Numerator Statement

Patient Population: The analysis population consists of patients aged 18 years or older who undergo isolated CABG, isolated AVR, AVR+CABG, isolated MVRR, and MVRR+CABG.

S.8. Denominator Exclusions: Measure exclusions: Individual surgeons who do not meet the minimum case requirement (i.e., at least 100 eligible cases during the 3-year measurement window) will not receive a score for each domain and an overall composite score.

De.1. Measure Type: Composite

S.17. Data Source: Registry Data

S.20. Level of Analysis: Clinician : Individual

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.

Evidence Summary of prior review in 2016

- This composite measure includes five major procedures (isolated CABG, isolated AVR, AVR+CABG, MVRR, MVRR+CABG) and 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 impact mortality and each of the five morbidities.
- The developer provided references that address operative mortality and morbidity dating from the 1990's 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:

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) → Relationship between outcome and healthcare action (Box 2) → Pass

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

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

Maintenance measures – increased emphasis on gap and variation

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

- The developer provided composite measure results for patients undergoing cardiac surgery during a three-year period, January 2017-December 2019. The developer included surgeons with at least 10 eligible records during the study period in the hierarchical model for estimating composite scores and noted that while surgeons with 10 eligible cases are included in the hierarchical model procedure, composite scores will typically only be reported by the STS for surgeons with at least 100 cases during a three-year time period. The developer did not provide performance gap information for the individual component measures.

  - The developer reports that 9.52% of surgeons with >100 cases (n = 1,841 surgeons with 584,571 operations) have lower than expected performance on the measure based on 98% Bayesian credible interval. In comparison, 9.51% of surgeons with >10 cases (n = 2,098 surgeons with 600,207 operations) have lower than expected performance.

Disparities

- The measure provides information about the performance of surgeons who participate in the STS database. The developer states that there is not a simple way to generate data stratified by patient characteristics at the composite level thus has not presented disparities data. Disparities data is a requirement for maintenance measures.
- In absence of or limited disparities data, NQF requires the developer provide a summary of data from the literature that addresses disparities in care. The developer did not provide a summary of data from the literature that addresses disparities in care either.

Questions for the Committee:

- *Is there a gap in care that warrants a national performance measure?*
- *Are you aware of evidence that disparities exist in this area of healthcare?*

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

RATIONALE: Data or literature summary that addresses disparities in care is not provided.

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 approach to development of the measure, including decision logic and results of testing (with STS registry data) used to combine the data into the respective domains (risk-adjusted operative mortality and risk-adjusted major morbidity) and then to combine domain scores into a single composite measure, is presented and described in detail (see Appendix Shahian et al) in a paper that addresses composite measure scoring and provider rating.
- The developer noted that this measure is based on a combination of risk-adjusted mortality and risk-adjusted major complications. To assess overall quality, the composite comprises two domains:
  - Domain 1 is risk-adjusted operative mortality (before hospital discharge or within 30 days of operation) for isolated coronary artery bypass grafting (CABG), isolated aortic valve replacement (AVR), AVR+CABG, isolated mitral valve repair or replacement (MVRR), and MVRR+CABG**.** This domain is calculated as a single measure.
  - Domain 2 is risk-adjusted major morbidity, which is an “any or none” measure of the following 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.
  - The developer states that the domains are rescaled by their respective standard deviation across surgeons and then assigned equal weighting to the rescaled rates. Using standard deviations derived from the data, the final composite measure is 0.81 x (1 minus risk-standardized mortality rate) + 0.19 x (1 minus risk-standardized complication rate).

- 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 results from five of the most frequently performed procedures and risk-adjusted occurrence of any of the five major complications, this composite provides a more comprehensive quality assessment that should help surgeons identify potential areas for improvement. By aggregating the surgeries and rates, the composite yields a more comprehensive view of surgeon performance, which may be more useful for accountability purposes.

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.

no info

Evidence continues to support.

They report no new evidence. I suspect you can find additional evidence, although I expect it will support the measure.

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?

disparities not discussed in measure

Continued lower than expected performance indicates room for improvement. No info on disparities.

Since this is a composite score for individual surgeons, difficult to look at disparities of care. However, hospital data per surgeon may give a glimpse to these issues. Likely addressed by a separate measure.

The numerator statement says that surgeons with >100 eligible cases will get domain and composite scores. So surgeons with <100 cases are not included? The description of the gap in quality includes surgeons with at least 10 cases. Why include low volume surgeons if they are excluded from the measure. 9.52% of surgeons with >100 cases (n = 1,841 surgeons with 584,571 operations) have lower than expected performance on the measure based on 98% Bayesian credible interval. What is the distribution of scores in this group? For surgeons with <=100 cases, the median is 95.4%. The minimum is 88%. The 10th percentile is 93%. Disparities data is a requirement for maintenance measures but it was not provided.

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?

no concerns

Quality construct and rationale seem appropriate.

Quality construct is logical and rational for weighing each factor explained.

The rationale for the composite is good. I am still unclear how much variation exists in each component (not just % in star categories).

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

No concerns with reliability.

Data elements are clearly defined. Appears reliable.

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

No concerns with reliability.

None

The data for reliability testing is almost 10 years old. Why wasn't this updated? Is there a NQF expectation that empirical testing be updated? The approach for calculating reliability is refreshingly thoughtful. My only small technical question is the choice of the squared true-to-observed correlation. Correlations just estimate the strength of linear association, not absolute agreement. In this context, agreement is important. Would a squared ICC[2, k] address this concern? Relatedly, it would be good to include numbers on the x and y-axis for the scatterplot. Also, the composite score is transformed into a star rating, but the reliability of that scoring system has not been determined. How stable are the star ratings. You could determine this will a split sample method or Bayesian methods.

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

no concerns

Unclear if all validity testing completed and submitted.

More complicated to assess due to individualized composite scores. Risk adjustment is a concern.

In fact, some 1-star surgeons might have better performance than some 2-star surgeons, but just tighter CIs.

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 concern

No issues.

No concerns.

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?

missing data does constitute threat to validity if low submission numbers

No major concerns.

It can. May be issue for individual surgeons if they don't meet the numbers (100 in 3 years) but would benefit from assessment.

9.52% of surgeons with >100 cases (n = 1,841 surgeons with 584,571 operations) have lower than expected performance on the measure based on 98% Bayesian credible interval. What is the distribution of scores in this group?

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 concern

No issues to composite construction.

appears well constructed

The added value of the components is unclear. A table summarizing the distributions of component performance (not just star ratings) would be helpful.

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 developer states that required data elements are generated or collected and used by healthcare personnel during provision of care. They are then

abstracted from a record by someone other than the person obtaining the original information (e.g., chart abstraction). Some of the elements are available in EHRs or from other electronic sources.

- Per the developer, the data elements in the 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 database has more than 1,000 participants. Local availability of data elements will vary from full EHR capability to no availability; however, all data elements are submitted to the STS database in electronic format following a standard set of data specifications.
- STS Adult Cardiac Surgery Database participants (single or group of surgeons) pay annual participant fees of $3,500 if majority of surgeons in the group are STS members and $4,750 if the majority are not STS members. In addition, there is a fee of $150 per member and $350 per non-member for surgeons listed on the database's Participation Agreement. There are no additional costs for data collection specific to the measure.

Questions for the Committee:

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

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 concerns

No issues or concerns with feasibility.

Elements obtained from STS Adult Cardiac Surgery Database. They are available electronically and are consistent.

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**

- This measure was initially endorsed in 2017. It is not currently in use in an accountability program however the developer has provided a highly credible path to public reporting, possibly as soon as this year. Concerns regarding the confidentiality and formatting of surgeon-level results delayed distribution of confidential surgeon-level feedback reports until January 2020. Providing a private review period of measure results prior to public reporting is a best practice. The developer has a strong record of publicly reporting measure results.

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. The developer noted 2,098 surgeons met the completeness and minimum procedure thresholds, 1,841 of whom performed at least 100 eligible cases within the three-year measurement period. Of this subset of surgeons, approximately 400 opted in for receipt of their confidential, surgeon-level performance results in January 2020. The report includes overall results, results by domain, benchmarks, and information on how to interpret the results.
2. The developer states that adult cardiac surgeons from across the U.S. and Canada who comprise the STS Adult Cardiac Surgery Database and Quality Measurement Task Forces meet periodically to discuss the surgeon-level and participant reports and to consider potential enhancements to the ACSD. Additions/clarifications to the data collection form and to the content/format of the individual surgeon reports and participant reports are discussed and implemented as appropriate.
3. The developer did not provide information on how this feedback has been considered when changes are incorporated into the measure

Additional Feedback:

- N/A

***Questions for the Committee*:**

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

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

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

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

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

Improvement results

- The developers state that they are unable to provide performance trends as performance data on this measure was only first distributed the consenting surgeons in January 2020.
- As a proxy for trend data on this measure, the developer provides 10 years of star rating trends for the five procedures aggregated within the composite. There is a general trend of reduction in participants receiving one or three stars and an increase in participants receiving two stars, which the developer states is consistent with their performance improvement goal of reducing variation.

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

Unexpected findings (positive or negative) during implementation

- The developer reports no unexpected findings during implementation.

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.

Additional Feedback:

- N/A

Questions for the Committee*:*

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

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

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

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

no concerns

Planned public reported as part of accountability program. Feedback provided and received for the measure, no clear indication of incorporation in measure development.

Data is not available to the public. For those meeting the required 100 procedures in 3 years, data has been presented to the surgeons and feedback obtained from them directly.

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.

no concerns

No usability issues.

Gaming and risk aversion controlled via audit and risk adjustment. Benefits of performance improvement still present, although results first given to use in 2020, so no trends available.

The composite has overall high performance.

Criterion 5: Related and Competing Measures

Related or competing measures

- The developers identified the following related 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
  - NQF #3032 Mitral Valve Repair/Replacement + CABG Composite Score

Harmonization

- The developer stated that the measure specifications have been harmonized to the extent possible.

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?

no concerns, harmonized to the extent possible

Multiple related measures, no concerns with harmonization.

Several related measures, including 2 reviewed in this cycle. They are harmonized as per developers.

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 Updates to Measure Testing Document Section 1b.4

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.

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 any of the 2018 risk adjustment models (see other sections for details of covariate adjustment – we used the most recent 2018 CABG, valve and 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.

Measures	Mortality: Adjusted OR (95% CI)	Mortality: p-value	Major Morbidity: Adjusted OR (95% CI)	Major Morbidity: p-value
Insurance status among patients age>=65	*	*	*	*
Medicare without Medicaid/Commercial-HMO	(ref)	*	(ref)	*
Medicare Medicaid dual eligible	0.95 (0.87, 1.03)	0.2178	1.05 (1.00, 1.09)	0.0537
Medicare Commercial-HMO without Medicaid	0.93 (0.89, 0.97)	0.0003	0.97 (0.95, 0.99)	0.0095

Measures	Mortality: Adjusted OR (95% CI)	Mortality: p-value	Major Morbidity: Adjusted OR (95% CI)	Major Morbidity: p-value
Commercial-HMO without Medicare	0.97 (0.90, 1.05)	0.448	1.00 (0.96, 1.04)	0.9403
Insurance status among patients age<65	*	*	*	*
Commercial-HMO without Medicare/Medicaid	(ref)	*	(ref)	*
Medicare or Medicaid	1.08 (1.01, 1.17)	0.0332	1.16 (1.12, 1.19)	<.0001
None/Self Paid	1.10 (0.98, 1.22)	0.099	1.08 (1.03, 1.13)	0.0022
Other	1.11 (0.96, 1.28)	0.151	1.03 (0.96, 1.09)	0.4283
Black Race	1.01 (0.95, 1.07)	0.8042	1.18 (1.15, 1.22)	<.0001
Hispanic ethnicity	1.00 (0.94, 1.07)	0.9194	1.01 (0.97, 1.04)	0.6444

*cell intentionally left blank

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.

Measure Number: 3030

Measure Title: Individual Surgeon Composite Measure for Adult Cardiac Surgery

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_3030" 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 is able to consistently implement and calculate the measure.

RELIABILITY: TESTING

Submission document: "MIF_3030" 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 composite-score-level signal-to-noise analysis. They utilized a Bayesian approach to generate possible values for each surgeon'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 100-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 0.77 (95% PrI 0.75 – 0.79) for 10 index cases to 0.82 (95% PrI 0.81 – 0.84) for 200 cases. At the planned public reporting threshold of 100 index cases, the reliability is 0.81 (95% PrI 0.79 – 0.82).

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 if score-level testing has been conducted)

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

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

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

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

    - 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 100 cases over 3 years (July 2011 – June 2014), around 71% of surgeons received 2 stars, and the remaining surgeons received either 1 or 3 stars.
    - 1,413 (71.5%) performed as expected (risk adjusted mortality, 2.5% and risk-adjusted morbidity, 14.2%);
    - 189 (9.6%) had lower-than-expected performance (risk adjusted mortality, 4.2% and risk-adjusted morbidity, 22.6%); and
    - 374 (18.9%) had higher-than-expected performance (risk adjusted mortality 1.2% and risk-adjusted morbidity, 8.8%)

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**

- The developer indicates they calculate a risk score for operative mortality and for major complications for each patient and use these patient-level scores to adjust for case mix. The scores were calculated using existing and modified risk models from the measures on which this measure is based. Calculating a risk score using this method limited the number of baseline covariates to a feasible number.
- The developer states that they validated this risk approach by performing sensitivity analyses comparing each surgeon’s risk-adjusted mortality and complication rates in models adjusting for 41 and 47 individual covariates with models adjusting for a single composite risk score.
- The developer provides odds ratios for its modified risk models.
- The developer reports the following c-statistics:

  - For modified MVRR model:
    - 0.746 for morbidity
    - 0.807 for mortality
  - Modified MVRR + CABG model:
    - 0.708 morbidity
    - 0.738 for mortality
- The developer interprets these results to indicate that the risk models are well calibrated and have 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. Using data from July 2011 – June 2014, the surgeons were divided into three groups as follows:
      - Surgeons were labeled as having higher-than-expected performance if the 98% credible interval surrounding a surgeon's composite score fell entirely above the overall STS average composite score.
      - Surgeons were labeled as having lower-than-expected performance if the 98% credible interval surrounding a surgeon's composite score fell entirely below the overall STS average composite score.
      - Surgeons 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 surgeons.
    - Demonstrating a relationship between performance on the overall composite and the composite domains may not be a valid assessment of composite score validity.

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

    Submission document: Testing attachment, section 2b2.3

    - The developers reported that compared to surgeons receiving 1 star, those with 3 stars had lower risk-adjusted mortality (1.2% vs. 4.2%) and lower risk-adjusted morbidity (8.8% vs. 22.6%) during July 2011 – June 2014. Thus, differences in performance were clinically meaningful as well as statistically significant. STS surgeons 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.

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 0.73 for mortality domain versus overall composite measure and 0.92 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 overall composite score was calculated as a weighted sum of (1 minus risk-adjusted mortality rate) and (1 minus risk-adjusted major morbidity rate). Mortality and morbidity rates were weighted

inversely by their respective standard deviations across surgeons. Standard deviations derived from the data were used to define the final composite measure as 0.81 × (1 minus risk-standardized mortality rate) + 0.19 × (1 minus risk-standardized complication rate).

- 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 surgeon's risk-adjusted mortality rate has the same impact on the overall score as a 4.3 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.**

Brief Measure Information

NQF #: 3030

Corresponding Measures:

De.2. Measure Title: STS Individual Surgeon Composite Measure for Adult Cardiac Surgery

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

De.3. Brief Description of Measure: The STS Individual Surgeon Composite Measure for Adult Cardiac Surgery includes five major procedures (isolated CABG, isolated AVR, AVR+CABG, MVRR, MVRR+CABG) and comprises the following two domains:

Domain 1 – Risk-Adjusted Operative Mortality

Operative mortality is defined as death before hospital discharge or within 30 days of the operation.

Domain 2 – Risk-Adjusted 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.

All measures are based on audited clinical data collected in the STS Adult Cardiac Surgery Database. Individual surgeons with at least 100 eligible cases during the 3-year measurement window will receive a score for each domain and an overall composite score. In addition to calculating composite score point estimates with credible intervals, surgeons will be assigned 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 in detail this multiprocedural, multidimensional composite measure.

The STS Individual Surgeon Composite Measure for Adult Cardiac Surgery includes five major procedures, i.e., isolated coronary artery bypass grafting (CABG), isolated aortic valve replacement (AVR), AVR+CABG, isolated mitral valve repair or replacement (MVRR), and MVRR+CABG, and comprises the following two domains:

Domain 1 – Risk-Adjusted Operative Mortality

Operative mortality is defined as death before hospital discharge or within 30 days of the operation.

Domain 2 – Risk-Adjusted 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

Individual surgeons with at least 100 eligible cases during the 3-year measurement window will receive a score for each domain and an overall composite score. In addition to calculating composite score point estimates with credible intervals, surgeons will be assigned 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 undergo isolated CABG, isolated AVR, AVR+CABG, isolated MVRR, and MVRR+CABG.

Time Window: 3 years

By including composite performance scores for a portfolio of five procedures that account for nearly 80% of a typical STS Adult Cardiac Surgery Database participant surgeon's clinical activity, this metric provides a more balanced and comprehensive perspective than focusing on just one procedure or one end point. Recognizing that surgeons' practices vary, each surgeon's composite performance is implicitly "weighted" by the proportion of each type of procedure he or she performs. For instance, the results of surgeons who primarily perform mitral procedures are affected most by their mitral surgery results. This approach is especially relevant for surgeons with highly specialized practices who may do relatively few isolated CABG procedures and whose performance would thus be difficult to assess using a CABG measure only. Finally, performance on each of these procedures is estimated using risk models specific to those procedures, in most cases the exact or slightly modified versions of previously published models (references provided below).

Final Composite Score:

The overall composite score was calculated as a weighted sum of (1 minus risk-adjusted mortality rate) and (1 minus risk-adjusted major morbidity rate). Mortality and morbidity rates were weighted inversely by their respective standard deviations across surgeons. This procedure is equivalent to first rescaling mortality and morbidity rates by their respective standard deviations across surgeons and then assigning equal weighting to the rescaled mortality rate and rescaled morbidity rate. Standard deviations derived from the data were used to define the final composite measure as 0.81 x (1 minus risk-standardized mortality rate) + 0.19 x (1 minus risk-standardized complication rate).

Details regarding the current STS adult cardiac surgery risk models can be found in the following manuscripts:

- Shahian DM, O´Brien SM, Filardo G, Ferraris VA, et al. The Society of Thoracic Surgeons 2008 cardiac surgery risk models: part 1--coronary artery bypass grafting surgery. Ann Thorac Surg. 2009 Jul;88(1 Suppl):S2-22.
- O'Brien SM, Shahian DM, Filardo G, et al. The Society of Thoracic Surgeons 2008 cardiac surgery risk models: part 2—isolated valve surgery. Ann Thorac Surg 2009;88(1 Suppl):S23–42.
- 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.

Additional details regarding the Individual Surgeon Composite Measure for Adult Cardiac Surgery are provided in the attached manuscript:

Shahian DM, He X, Jacobs JP, Kurlansky PA, Badhwar V, Cleveland JC Jr, Fazzalari FL, Filardo G, Normand SL, Furnary AP, Magee MJ, Rankin JS, Welke KF, Han J, O´Brien SM. The Society of Thoracic Surgeons Composite

Measure of Individual Surgeon Performance for Adult Cardiac Surgery: A Report of The Society of Thoracic Surgeons Quality Measurement Task Force. Ann Thorac Surg. 2015;100:1315-25.

S.6. Denominator Statement: See response in S.4. Numerator Statement

Patient Population: The analysis population consists of patients aged 18 years or older who undergo isolated CABG, isolated AVR, AVR+CABG, isolated MVRR, and MVRR+CABG.See response in S.4. Numerator Statement

Patient Population: The analysis population consists of patients aged 18 years or older who undergo isolated CABG, isolated AVR, AVR+CABG, isolated MVRR, and MVRR+CABG.

S.8. Denominator Exclusions: Measure exclusions: Individual surgeons who do not meet the minimum case requirement (i.e., at least 100 eligible cases during the 3-year measurement window) will not receive a score for each domain and an overall composite score.

De.1. Measure Type: Composite

S.17. Data Source: Registry Data

S.20. Level of Analysis: Clinician : Individual

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.***

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-3030-Surg_Comp_Adult_Cardiac_Surg-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*)**:** 3030

Measure Title: Individual Surgeon Composite Measure for Adult Cardiac Surgery

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**

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

Mortality likely is the single most important negative outcome associated with a surgical procedure. Operative mortality, defined as death before hospital discharge or within 30 days of the operation, should include nearly all deaths that occur as a direct result of the surgery or an immediate postoperative complication. 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. Preoperative patient selection, surgical timing post coronary event, intraoperative conduct of the case, and many aspects to postoperative care have all been shown to have significant impact on the operative mortality over the last few decades. The published literature (list provided below) on each major procedure included in this composite measure is full of examples of services/care processes that impact operative mortality.

Major Morbidity

- Surgical re-exploration for bleeding remains a known complication following cardiac surgery. The literature documents that bleeding following coronary artery bypass surgery confers greater ICU stay and therefore greater resource consumption. It remains unknown and controversial whether long-term outcomes are worse for the isolated re-exploration for bleeding patients. However, Hein documents that patients with ICU stay > 3 days (with bleeding as multivariate risk factor for this outcome), have a long-term survival which is inferior to patients with ICU stay < 3 days. The patient consequences of this complication relates to the physiological stress of facing another operation and receiving blood products.
- A devastating complication of cardiac surgery is deep sternal wound infection. Patients with deep sternal wound infection require multiple surgeries to clear the infection, 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.

- Prolonged ventilation has been shown to substantially increase length of stay, the costs 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 protocols for extubation 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 to increase the implementation of evidence based care exist. Cardiac surgery programs with high implementation have lower than average rates of prolonged ventilation and significantly lower rates of adverse events.
- 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 treatment of this high-risk group will improve the long-term survival of our 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.
- Postoperative stroke/CVA produces significant short- and long-term often devastating effects to 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 perfusion, glycemic control, avoidance of atrial fibrillation, anticoagulation protocols, etc. Many opportunities exist to decrease stroke rates by increasing implementation of evidence based strategies.

References – Operative Mortality

- Ferguson TB, Hammill BG, et al. A decade of change—risk profiles and outcomes for isolated coronary artery bypass grafting procedures, 1990-1999: a report from the STS National Database Committee and the Duke Clinical Research Institute. *Ann Thorac Surg.* 2002;73(2):480-489; discussion 489-490.
- Grover FL, Shroyer AL, et al. A decade's experience with quality improvement in cardiac surgery using the Veterans Affairs and Society of Thoracic Surgery national databases. *Ann Thorac Surg.*2001; 234(4):464-472; discussion 472-474.
- Hogue CW, Barzilai B, et al. Sex differences in neurologic outcomes and mortality after cardiac surgery: A Society of Thoracic Surgeons National Database report. *Circulation.*2001;03:2133-2137.
- 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;75:1856-1865.
- Williams ML, Muhlbaier LH, Schroder JN, et. al. Risk-adjusted short- and long-term outcomes for on-pump versus off-pump coronary artery bypass surgery. Circulation. 2005 Aug 30;112(9 Suppl):I366-70.
- Shroyer AL, Grover FL, Hattler B, et. al. On-pump versus off-pump coronary artery bypass surgery. N Engl J Med. 2009 Nov 5;361(19):1827-37.
- Hannan EL, Wu C, Smith CR, et. al. Off-pump versus on-pump coronary artery bypass graft surgery: differences in short-term outcomes and in long-term mortality and need for subsequent revascularization. Circulation. 2007 Sep 4;116(10):1145-52. Epub 2007 Aug 20.
- ElBardissi AW, Aranki SF, Sheng S, et al. Trends in isolated coronary artery bypass grafting: an analysis of the Society of Thoracic Surgeons adult cardiac surgery database. J Thorac Cardiovasc Surg. 2012 Feb;143(2):273-81.

- Rangrass G, Ghaferi AA, Dimick JB. Explaining Racial Disparities in Outcomes After Cardiac Surgery: The Role of Hospital Quality. JAMA Surg. 2014;149(3):223-7
- Birkmeyer NJ, 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. JACC. 37:3:885-892.
- 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;76:1131-1137.
- Mihaljevic T, Nowicki ER, Rajeswaran J, et. al. Survival after valve replacement for aortic stenosis: implications for decision making. J Thorac Cardiovasc Surg. 2008 Jun;135(6):1270-8; discussion 1278-9.
- Tabata M, Umakanthan R, Cohn LH, et. al. Early and late outcomes of 1000 minimally invasive aortic valve operations. Eur J Cardiothorac Surg. 2008;33(4):537-41.
- Chaliki HP, Mohty D, Avierinos JF, et. al. Outcomes after aortic valve replacement in patients with severe aortic regurgitation and markedly reduced left ventricular function. Circulation. 2002 Nov 19;106(21):2687-93.
- Brennan JM, Holmes DR, Sherwood MW, Edwards FH, et al. The Association of Transcatheter Aortic Valve Replacement Availability and Hospital Aortic Valve Replacement Volume and Mortality in the United States. Ann Thorac Surg. 2014 Dec;98(6):2016-22.
- Thourani VH, Suri RM, et al. Contemporary real-world outcomes of surgical aortic valve replacement in 141,905 low-risk, intermediate-risk, and high-risk patients. Ann Thorac Surg. 2015 Jan;99(1):55-61.
- Chikwe J, Croft LB, Goldstone AB, Castillo JG, Rahmanian PB, Adams DH, et al . Comparison of the results of aortic valve replacement with or without concomitant coronary artery bypass grafting in patients with left ventricular ejection fraction ≤30%versus patients with ejection fraction > 30%. Am J Cardiol. 2009;104:1717-21.
- Li Z, Anderson I, Amsterdam EA, Young N, Parker J and Armstrong EJ. Effect of coronary artery disease extent on contemporary outcomes of combined aortic valve replacement and coronary artery bypass graft surgery. Ann Thor Surg 2013;96:2075-82.
- Kobayashi J. Changing strategy for aortic stenosis with coronary artery disease by transcatheter aortic valve implantation. Gen Thorac Cardiovas Surg 2013;61:663-68.
- Beach JM Mihaljevic T, Svensson LG, Rajeswaran J, Marwich T, Griffin B, Johnston DR, Sabik III JF and Blackstone EJ. Coronary artery disease and outcomes of aortic valve replacement for severe aortic stenosis. J Am Coll Cardiol 2013;61:837-48.
- Fukui T, Bando K, Tanaka S, Uchimuro T, Tabata M and Takanashi S. Early and mid-term outcomes of combined aortic valve replacement and coronary artery bypass grafting in elderly patients. Eur J of Cardio-Thorac Surg 2014;45:335-40.
- Mehta RH, Eagle KA, et al. Influence of age on outcomes in patients undergoing mitral valve replacement. Ann Thorac Surg. 2002;74:1459-1467.
- Dayan V, Soca G, et al. Similar survival after mitral valve replacement or repair for ischemic mitral regurgitation: a meta-analysis. Ann Thorac Surg. 2014 Mar;97(3):758-65.
- Kaneko T, Aranki S, et al. Mechanical versus bioprosthetic mitral valve replacement in patients <65 years old. J Thorac Cardiovasc Surg. 2014 Jan;147(1):117-26.
- Iribarne A, Russo MJ, Easterwood R et al. Minimally invasive versus sternotomy approach for mitral valve surgery: a propensity analysis. *Ann Thorac Surg.* 2010;90:1471–1477

- LaPar DJ, Hennessy S, Fonner E, et al. Does urgent or emergent status influence choice in mitral valve operations? An analysis of outcomes from the Virginia Cardiac Surgery Quality Initiative. 2010;90:153-60
- Umakanthan R, Petracek MR, Leacche M et al, Minimally invasive right lateral thoracotomy without aortic cross-clamping: an attractive alternative to repeat sternotomy for reoperative mitral valve surger;y. J Heart Valve Dis. 2010;19:236-43
- Vassileva CM, McNeely C, Spertus J, Markwell S, Hazelrigg S. Hospital volume, mitral repair rates, and mortality in mitral valve surgery in the elderly: An analysis of US hospitals treating Medicare fee-for-service patients. J Thorac Cardiovasc Surg. 2014pii: S0022-5223(14)01290-2
- Chatterjee S, Rankin JS, Gammie JS, et al. Isolated mitral valve surgery risk in 77,836 patients from the Society of Thoracic Surgeons database. Ann Thorac Surg. 2013;96:1587-94
- LaPar DJ, Ailawadi G, Isbell JM, et al. Virginia Cardiac Surgery Quality Initiative. Mitral valve repair rates correlate with surgeon and institutional experience. J Thorac Cardiovasc Surg. 2014;148:995-1003
- 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. 2010 Nov 1
- 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. 2010 Aug 18.
- 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 Nov; 90(5):1479-85
- Acker MA, Parides MK, Perrault LP et al (members of Cardiothoracic Surgical Trials Network). Mitral-valve repair versus replacement for severe ischemic mitral regurgitation. N Engl J Med 2014; 370:23-32

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;75:1856-1865.
- Welke KF, Ferguson TB, Coombs LP, et al. Validity of the Society of Thoracic Surgeons National Adult Cardiac Surgery Database. *Ann Thorac Surg.* 2004;77:1137-1139.
- 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.
- Karthik S, Grayson AD, McCarron EE, Pullan DM, Desmond MJ. Reexploration for bleeding after coronary artery bypass surgery: risk factors, outcomes, and the effect of time delay. *Ann Thor Surg* 2004;78:527-34.
- Stamou SC, Camp SL, Stiegel RM, et al. Quality improvement program decreases mortality after cardiac surgery. *J Thorac Cardiovasc Surg* 2008;136:494-499.
- 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.
- Speir AM, Kasirajan V, BarnettSD, Fonner E. Additive costs of postoperative complications for isolated coronary artery bypass grafting patients in Virginia. Ann Thorac Surg 2009;88:40-46.
- Olsen MA, Lock-Buckley P, et al. The risk factors for deep and superficial chest surgical site infections after coronary artery bypass graft surgery are different. J Thorac Cardiovasc Surg. 2002;124:136-145.

- Shahian DM, Edwards FH, Ferraris VA, et al. Quality measurement in adult cardiac surgery; part1 – Conceptual framework and measure selection. Ann Thorac Surg 2007;83:S3-S12.
- Shahian DM, O'Brien SM, Filardo G, et al. The Society of Thoracic Surgeons 2008 cardiac surgery risk models: part 1 – coronary artery bypass grafting surgery. Ann Thorac Surg 2009;88(1 Suppl):S2-S22.
- Shahian DM, Edwards FH, Jacobs JP, et al. Public reporting of cardiac surgery performance: part 2 – implementation. Ann Thorac Surg 2011;92:S12-S23.
- Trick WE, Scheckler WE, et al. Modifiable risk factors associated with deep sternal site infection after coronary artery bypass grafting. J Thorac Cardiovasc Surg. 2000;119:108-114.
- 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 APL, 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 – 24.
- 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;144:931-37.
- 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, MatsuuraA, Takemura H, et al. Implementation of bundled interventions greatly decreases deep sternal wound infection following cardiovascular surgery. J Thorac Cardiovasc Surg 2014;148:2381-88.
- Matros E, Aranki, SF, Bayer LR, et al. Reduction in incidence of deep sternal wound infections: random or real? J Thorac Cardiovasc Surg 2010;139:680-85.
- Bardell T, Legare JF, Buth KJ, et al. ICU readmission after cardiac surgery. Eur J Cardiothorac Surg. 2003;23(3):354-359.
- Meade MO, Guyatt G, Butler R, et al. Trials comparing early vs late extubation following cardiovascular surgery. Chest. 2001:120(6 Suppl):445S-453S.
- 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(3):225-233.
- Engel AM, McDonough S, Smith JM. Does an obese body mass index affect hospital outcomes after coronary artery bypass graft surgery? Ann Thorac Surg. 2009 Dec;88(6):1793-800.
- Brown PP, Kugelmass AD, Cohen DJ, Reynolds MR, Culler SD, Dee AD, Simon AW. The frequency and cost of complications associated with coronary artery bypass grafting surgery: results from the United States Medicare program. Ann Thorac Surg. 2008 Jun;85(6):1980-6. PubMed PMID: 18498806.
- 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.
- Hesham Z, Saleha HZ, Shawb M, et al. Outcomes and predictors of prolonged ventilation in patients undergoing elective coronary surgery. Interact Cardiovasc Thorac Surg 2012;15:51–6.
- Jacobs JP, He X, O'Brien SM, Welke KF, Filardo G, Han JM, Ferraris VA, Prager RL, Shahian DM.. Variation in Ventilation Time after Coronary Artery Bypass Grafting: An Analysis from The Society of Thoracic Surgeons Adult Cardiac Surgery Database. Ann Thorac Surg. 2013 Sep;96(3):757-62.

- 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; 26:687-697.
- 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(3):906-912.
- Chertow GM, Levy EM, Hammermeister KE, et al. Independent association between acute renal failure and mortality following cardiac surgery. Am J Med. 1998;104(4):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(5):1158-1162.
- Cooper WA, O'Brien SM, Thourani VH, Guyton RA, Bridges CR, Szczech LA, Petersen R, Peterson ED. Impact of renal dysfunction on outcomes of coronary artery bypass surgery: results from the Society of Thoracic Surgeon's National Adult Cardiac Database. Circulation. 2006;113:1063-1070.
- Gallagher S, Jones DA, Lovell MJ, Hassan S, Wragg A, Kapur A, Uppal R, Yaqoob MM. The impact of acute kidney injury on midterm outcomes after coronary artery bypass graft surgery: a matched propensity score analysis. J Thorac Cardiovasc Surg 2104;147:989-995.
- Haase M, Haase-Fielitz A, Bellomo R, Devarajan P, Story D, Matalanis G, Reade MC, Bagshaw SM, Seevanayagam N, Seevanayagam S, Doolan L, Buxton B, Dragun D. Sodium bicarbonate to prevent increases in serum creatinine after cardiac surgery: a pilot double-blind, randomized trial. Crit Care Ned 2009;37:39-47.
- Hillis GS, Croal BL, Buchan KG, El-Shafei H, Gibson G, Jeffrey RR, Millar CGM, Prescott GJ, Cuthbertson BH. Renal function and outcome from coronary artery bypass grafting: impact on mortality after 2.3-year follow up. Circulation. 2006;113:1056-1062.
- Hillis LD, Smith PK, Anderson JL, Bittl JA, Bridges CR, Byrne JG, Cigarroa JE, DiSesa VJ, Hiratzka LF, Hutter AM, Jessen ME, Keeley EC, Lahey SJ, Lange RA, London MJ, Mack MJ, Patel MR, Puskas JD, Sabik JF, Selnes O, Shahian DM, Trost JC, Winniford MD. 2011 ACC/AHA guideline for coronary artery bypass graft surgery: executive summary. A report of the American College of Cardiology Foundation/American Heart Association Task Force on Practice Guidelines. Circulation. 2011;124:2610 -2642.
- Karthik S, Musleh G, Grayson AD, Keenan DJ, Hasan R, Pullan DM, Dihmis WC, Fabri BM. Effect of avoiding cardiopulmonary bypass in non-elective coronary artery bypass surgery: a propensity score analysis. Eur J Cardiothorac Surg. 2003;24(1):66-71.
- 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.
- Kuss O, von Salviati B, Borgermann J. Off-pump versus on-pump coronary artery bypass grafting: a systematic review and meta-analysis of propensity score analyses. J Thorac Cardiovasc Surg 2010;140:829-35.
- Lamy A, Devereaux PJ, Prabhakaran D, Taggart DP, Hu S, Paolasso E, Straka Z, Piegas LS, Akar AR, Jain AR, Noiseux N, Padmanabhan C, Bahamondes JC, Novick R, Vaijyanath P, Reddy S, Tao L, Olavegogeascoechea PA, Airan B, Sulling TA, Whitlock RP, Ou Y, Ng J, Chrolavicius S, Yusuf S for the CORONARY Investigators. Off-Pump or On-Pump Coronary-Artery Bypass Grafting at 30 Days. N Engl J Med 2012;366:1489-97.
- 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(3):194-203.

- Medalion B, Cohen H, C, Assali A, Vaknin Assa H, Farkash A, Snir E, Sharoni E, Biderman P, Milo G, Battler A, Kornowski R, Porat E. The effect of cardiac angiography timing, contrast media dose, and preoperative renal function on acute renal failure after coronary artery bypass grafting. J Thorac Cardiovasc Surg 2010;139:1539-44.
- 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 2103;95:513-519.
- Rosner MH, Okusa MD. Acute kidney injury associated with cardiac surgery. Clin J Am Soc Nephrol 2006;1:19-32.
- Seabra VF, Alobaidi S, Balk EM, Poon AH, Jaber BL. Off-pump coronary artery bypass surgery and acute kidney injury: a meta-analysis of randomized controlled trials. Clin J Am Soc Nephrol 2010;5:1734-1744.
- 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
- Shroyer AL, Grover FL, Hattler B, Collins JF, McDonald GO, Kozora E, Lucke JC, Baltz JH, Novitzky D, for the Veterans Affairs Randomized On/Off Bypass (ROOBY) Study Group. On-Pump versus Off-Pump Coronary-Artery Bypass Surgery. N Engl J Med 2009;361:1827-37.
- Stallwood MI, Grayson AD, Mills K, et al. Acute renal failure in coronary artery bypass surgery: independent effect of cardiopulmonary bypass. Ann Thorac Surg. 2004;77(3):968-972.
- 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(2):372-327; discussion 377.
- Afilalo J, Rasti M, Ohayon SM, Shimony A, Eisenberg MJ. Off-pump vs on-pump coronary bypass surgery: an updated meta-analysis and meta-regression of randomized trials. Eur Heart J. 2012; 33:1257-67
- 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(3):270-277.
- 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. Cocrane 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(1):57-64.
- Engelman DT, Cohn LH, Rizzo RJ. Incidence and predictors of TIAs and strokes following coronary artery bypass grafting: report and collective review. Heart Surg Forum. 1999;2(3):242-245.
- 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(2):201-208.
- Likosky DS, Leavitt BJ, Marrin CA, et al. Intra- and postoperative predictors of stroke after coronary artery bypass grafting. Ann Thorac Surg. 2003;76(2):428-434.
- Mangano DT. Aspirin and mortality from coronary bypass surgery. N Engl J Med. 2002; 347(17):1309-1317.
- Puskas JD, Winston AD, Wright CE, et al. Stroke after coronary artery operation: incidence, correlates, outcome and cost. Ann Thorac Surg. 2000:69(4):1053-1056.

- Brown PP, Kugelmass AD, Cohen DJ, Reynolds MR, Culler SD, Dee AD, Simon AW. The frequency and cost of complications associated with coronary artery bypass grafting surgery: results from the United States Medicare program. Ann Thorac Surg. 2008 Jun;85(6):1980-6. PubMed PMID: 18498806.
- Naylor AR. Does the risk of post-CABG stroke merit staged or synchronous reconstruction in patients with symptomatic or asymptomatic carotid disease? J Cardiovasc Surg (Torino). 2009 Feb;50(1):71-81.
- Bouchard D, Carrier M, Demers P, Cartier R, Pellerin M, Perrault LP, et al. Statin in combination with beta blocker therapy reduces postoperative stroke after coronary artery bypass graft surgery. Ann Thorac Surg. 2011:91(3) 654-9.
- Rosenberger P, Shernan SK, Loffler M, Shekar PS, Fox JA, Tuli JK, Nowak M and Eltzschig HK. The influence of epiaortic ultrasonography n intraoperative surgical management in 6051 cardiac surgical patients. Ann Thorac Surg. 2008; 85: 548-53.
- Mehaffey JH, Hawkins RB, Byler M, Charles EJ, Fonner C, Kron I, Quader M, Speir A, Rich J, Ailawadi G; Virginia Cardiac Services Quality Initiative. Cost of individual complications following coronary artery bypass grafting. J Thorac Cardiovasc Surg. 2018 Mar;155(3):875-882.e1. doi: 10.1016/j.jtcvs.2017.08.144. Epub 2017 Dec 14. PMID: 29248284.
- Alshaikh HN, Katz NM, Gani F, Nagarajan N, Canner JK, Kacker S, Najjar PA, Higgins RS, Schneider EB. Financial Impact of Acute Kidney Injury After Cardiac Operations in the United States. Ann Thorac Surg. 2018 Feb;105(2):469-475. doi: 10.1016/j.athoracsur.2017.10.053. Epub 2017 Dec 21. PMID: 29275828.
- Edwards FH, Ferraris VA, Kurlansky PA, Lobdell KW, He X, O'Brien SM, Furnary AP, Rankin JS, Vassileva CM, Fazzalari FL, Magee MJ, Badhwar V, Xian Y, Jacobs JP, Wyler von Ballmoos MC, Shahian DM. Failure to Rescue Rates After Coronary Artery Bypass Grafting: An Analysis From The Society of Thoracic Surgeons Adult Cardiac Surgery Database. Ann Thorac Surg. 2016 Aug;102(2):458-64. doi: 10.1016/j.athoracsur.2016.04.051. Epub 2016 Jun 22. PMID: 27344280.

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

Systematic Review	Evidence
Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	*

*cell intentionally left blank

____

1a.4 **OTHER SOURCE OF EVIDENCE**

If source of evidence is NOT from a clinical practice guideline, USPSTF, or systematic review, please describe the evidence on which you are basing the performance measure.

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 data for patients undergoing cardiac surgery during January 2017 - December 2019. Five major procedures were included: isolated CABG, isolated AVR, AVR + CABG, isolated mitral valve repair or replacement procedures, and mitral valve repair or replacement + CABG procedures. Initially, 2919 surgeons were identified using their National Provider Identifiers. Surgeons without a National Provider Identifier (e.g., from a foreign country) or with invalid National Provider Identifiers were excluded (49). Surgeons were also required to have reported at least one of any type of cardiac procedure during each of the three 12month periods (i.e., January-December 2017, 2018 and 2019). This was to ensure that the included surgeons had not just finished training or, conversely, had retired, and that they had actively participated in the most recent STS harvest. This requirement excluded 495 surgeons. From the remaining 2,375 surgeons, we included 2095 surgeons who met the annual completeness threshold of 98% of the operative mortality fields to assure accuracy of the operative mortality endpoint and had performed at least 10 major procedures during the 3-year period, both to facilitate statistical computations and because results would not be calculated or reported for surgeons with lower volumes than this. In the table below, we provide the number of measured entities (# surgeons), the number of eligible patient records (# operations), and the distribution of composite score estimates by percentiles. Surgeons with at least 10 eligible records during the study period were included in the hierarchical model for estimating composite scores. While surgeons with 10 eligible cases are included in the hierarchical model procedure, composite scores will typically only be reported by the STS for surgeons with at least 100 cases during a 3-year time period. Thus, we tabulate results for all eligible surgeons and the subset with at least 100 eligible cases. Please see appendix for the histogram that summarizes the distribution of scores across surgeons.

Stat	Surgeons with >=10 Eligible Cases	Surgeons with >=100 Eligible Cases
# Participant	2098	1841
# Operations	600207	584571
Mean	0.951	0.952
STD	0.01547	0.01508
IQR	0.0198	0.0193
0%	0.869	0.886
10%	0.931	0.932
20%	0.939	0.940
30%	0.945	0.946
40%	0.949	0.950
50%	0.953	0.954
60%	0.957	0.958
70%	0.960	0.961
80%	0.964	0.965
90%	0.969	0.969
100%	0.984	0.984

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.*

This composite measure gauges the performance of STS surgeons and is not a patient or operation level measure. We do not have a simple way to generate data stratified by patient characteristics at the composite level.

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.

Suitable for evaluating surgical performance of individual adult cardiac surgeons, the STS Individual Surgeon Composite Measure for Adult Cardiac Surgery is based on aggregate risk-adjusted morbidity and mortality for five common procedures, i.e., isolated coronary artery bypass grafting (CABG), isolated aortic valve replacement (AVR), AVR+CABG, isolated mitral valve repair or replacement (MVRR), and MVRR+CABG. 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. To assess overall quality, the composite comprises the following two domains:

Domain 1 – Risk-Adjusted Operative Mortality

Operative mortality is defined as death before hospital discharge or within 30 days of the operation.

Domain 2 – Risk-Adjusted 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.

This composite measure differs from the NQF-endorsed, program-level STS CABG Composite Score in that it does not include the two process measure domains (use of internal mammary artery in CABG and perioperative medications). This approach was necessary for computational reasons to efficiently combine the results from five procedures, most of which did not have comparable process measures available.

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 surgery procedures, but in an era when the average mortality rates for these procedures have declined to very low levels, differentiating performance based on mortality alone is difficult. Specifically, it fails to take into account the fact that not all operative survivors received equal quality care, e.g., patients who survive surgery but have a debilitating complication that may substantially impact long-term freedom from cardiac events.

In addition, with the development of this composite measure, STS addresses a number of major concerns that have previously been raised regarding surgeon-level metrics. It combines results from five of the most frequently performed cardiac surgical procedures, encompassing most of a typical adult cardiac surgeon's practice, as opposed to basing performance on just one or a few separate procedures. Furthermore, it provides a more comprehensive quality assessment and additional endpoints, as it includes risk-adjusted mortality and the risk-adjusted occurrence of any of five major complications. This measure will be useful to surgeons in identifying potential areas for improvement, and it has numerous advantages compared with existing surgeon metrics if used for accountability purposes.

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.

Mortality and morbidity rates were weighted inversely by their respective standard deviations across surgeons. This procedure is equivalent to first rescaling mortality and morbidity rates by their respective standard deviation across surgeons and then assigning equal weighting to the rescaled mortality rate and rescaled morbidity rate. Standard deviations derived from the data were used to define the final composite measure as 0.81 × (1 minus risk-standardized mortality rate) + 0.19 × (1 minus risk-standardized complication rate).

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):*

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 in detail this multiprocedural, multidimensional composite measure.

The STS Individual Surgeon Composite Measure for Adult Cardiac Surgery includes five major procedures, i.e., isolated coronary artery bypass grafting (CABG), isolated aortic valve replacement (AVR), AVR+CABG, isolated mitral valve repair or replacement (MVRR), and MVRR+CABG, and comprises the following two domains:

Domain 1 – Risk-Adjusted Operative Mortality

Operative mortality is defined as death before hospital discharge or within 30 days of the operation.

Domain 2 – Risk-Adjusted 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

Individual surgeons with at least 100 eligible cases during the 3-year measurement window will receive a score for each domain and an overall composite score. In addition to calculating composite score point estimates with credible intervals, surgeons will be assigned 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 undergo isolated CABG, isolated AVR, AVR+CABG, isolated MVRR, and MVRR+CABG.

Time Window: 3 years

By including composite performance scores for a portfolio of five procedures that account for nearly 80% of a typical STS Adult Cardiac Surgery Database participant surgeon's clinical activity, this metric provides a more balanced and comprehensive perspective than focusing on just one procedure or one end point. Recognizing that surgeons' practices vary, each surgeon's composite performance is implicitly "weighted" by the proportion of each type of procedure he or she performs. For instance, the results of surgeons who primarily perform mitral procedures are affected most by their mitral surgery results. This approach is especially relevant for surgeons with highly specialized practices who may do relatively few isolated CABG procedures and whose performance would thus be difficult to assess using a CABG measure only. Finally, performance on each of these procedures is estimated using risk models specific to those procedures, in most cases the exact or slightly modified versions of previously published models (references provided below).

Final Composite Score:

The overall composite score was calculated as a weighted sum of (1 minus risk-adjusted mortality rate) and (1 minus risk-adjusted major morbidity rate). Mortality and morbidity rates were weighted inversely by their respective standard deviations across surgeons. This procedure is equivalent to first rescaling mortality and morbidity rates by their respective standard deviations across surgeons and then assigning equal weighting to the rescaled mortality rate and rescaled morbidity rate. Standard deviations derived from the data were used to define the final composite measure as 0.81 x (1 minus risk-standardized mortality rate) + 0.19 x (1 minus risk-standardized complication rate).

Details regarding the current STS adult cardiac surgery risk models can be found in the following manuscripts:

- Shahian DM, O´Brien SM, Filardo G, Ferraris VA, et al. The Society of Thoracic Surgeons 2008 cardiac surgery risk models: part 1--coronary artery bypass grafting surgery. Ann Thorac Surg. 2009 Jul;88(1 Suppl):S2-22.
- O'Brien SM, Shahian DM, Filardo G, et al. The Society of Thoracic Surgeons 2008 cardiac surgery risk models: part 2—isolated valve surgery. Ann Thorac Surg 2009;88(1 Suppl):S23–42.
- 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.

Additional details regarding the Individual Surgeon Composite Measure for Adult Cardiac Surgery are provided in the attached manuscript:

Shahian DM, He X, Jacobs JP, Kurlansky PA, Badhwar V, Cleveland JC Jr, Fazzalari FL, Filardo G, Normand SL, Furnary AP, Magee MJ, Rankin JS, Welke KF, Han J, O´Brien SM. The Society of Thoracic Surgeons Composite Measure of Individual Surgeon Performance for Adult Cardiac Surgery: A Report of The Society of Thoracic Surgeons Quality Measurement Task Force. Ann Thorac Surg. 2015;100:1315-25.

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

Patient Population: The analysis population consists of patients aged 18 years or older who undergo isolated CABG, isolated AVR, AVR+CABG, isolated MVRR, and MVRR+CABG.See response in S.4. Numerator Statement

Patient Population: The analysis population consists of patients aged 18 years or older who undergo isolated CABG, isolated AVR, AVR+CABG, isolated MVRR, and MVRR+CABG.

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.6. Denominator Statement

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

Measure exclusions: Individual surgeons who do not meet the minimum case requirement (i.e., at least 100 eligible cases during the 3-year measurement window) will not receive a score for each domain and an overall composite score.

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 : Individual

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 and Appendix.

2. Validity – See attached Measure Testing Submission Form

3030_NQF_testing_v3.0-SurgeonComp-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*)**:** 3030
Composite Measure Title: Individual Surgeon Composite Measure for Adult Cardiac Surgery
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)

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

Measure Specified to Use Data From: **(*must be consistent with data sources entered in S.17*)**	**Measure Tested with Data From:**
☐ 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 data from 2286 surgeons for patients undergoing cardiac surgery during July 2011 – June 2014. Only surgeons 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 surgeon-specific denominators (number of eligible patients) and surgeon-specific mortality and morbidity rates.

Stat	N (Denominator)	% Mortality	% Morbidity
N	2286	2286	2286
Mean	272	2.5	14.8
STD	166	1.8	6.7
IQR	208	2.0	7.7
0%	10	0.0	0.0
10%	79	0.7	7.7
20%	132	1.1	9.5
30%	170	1.5	10.9
40%	209	1.9	12.3

Stat	N (Denominator)	% Mortality	% Morbidity
50%	250	2.2	13.8
60%	293	2.6	15.3
70%	336	3.1	17.0
80%	394	3.7	19.2
90%	478	4.8	23.0
100%	1435	15.4	64.8

1.6. How many and which patients were included in the testing and analysis (by level of analysis and data source)? (*identify the number and descriptive characteristics of patients included in the analysis (e.g., age, sex, race, diagnosis); if a sample was used, describe how patients were selected for inclusion in the sample*)
The study cohort included 621,489 patient operations performed by 2,286 surgeons during July 2011 – June 2014. The number of patients and their unadjusted outcomes are summarized by procedure type in the following table.

Procedure Group	Total: No. (%)	Mortality: No.	Mortality: Rate, %	Morbidity: No.	Morbidity: Rate, %
Isolated CABG	417,261 (67.1)	8,295	2.0	51,281	12.3
Isolated AVR	84,751 (13.6)	2,059	2.4	11,458	13.5
Isolated MVR	14,948 (2.4)	539	3.6	2,905	19.4
AVR + CABG	53,081 (8.5)	2,124	4.0	10,801	20.3
MVR + CABG	6,547 (1.1)	474	7.2	2,125	32.5
Isolated MV repair	30,347 (4.9)	339	1.1	2,953	9.7
MV repair + CABG	14,554 (2.3)	635	4.4	3,694	25.4
Total	621,489 (100.0)	14,465	2.3	85,217	13.7

AVR = aortic valve replacement

CABG = coronary artery bypass grafting

MV = mitral valve

MVR = mitral valve replacement

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

The methodology was developed and tested using data from 621,489 patients operations performed by 2,286 surgeons. To ensure adequate statistical precision, STS plans to report composite scores only for surgeons with at least 100 eligible cases during the 3-year measurement window. Thus, some of the analyses in this submission are limited to surgeons with at least 100 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 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.

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*)

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_j denote the true unknown composite measure value for the j-th of J surgeons. Before estimating reliability, the numeric value of θ_j was estimated for each surgeon 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_j 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 surgeon.

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 composite measure using 3 years of data in surgeons with at least 100 total cases was 0.81 (95% CrI, 0.79 to 0.82), as outlined in the Table below. For comparison, 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).

Threshold Number of Index Cases Over 3 Years	Surgeon Included (No.)	Patients Included (No.)	Reliability p^2 (95% PrI)
10	2,286	621,489	0.77 (0.75, 0.79)
25	2,234	620,586	0.78 (0.76, 0.80)
36	2,205	619,691	0.79 (0.77, 0.80)
50	2,165	617,976	0.79 (0.78, 0.81)
100	1,976	603,594	0.81 (0.79, 0.82)
150	1,737	573,491	0.81 (0.80, 0.83)
200	1,432	520,724	0.82 (0.81, 0.84)

[a] Number of surgeons and patients included for each threshold.

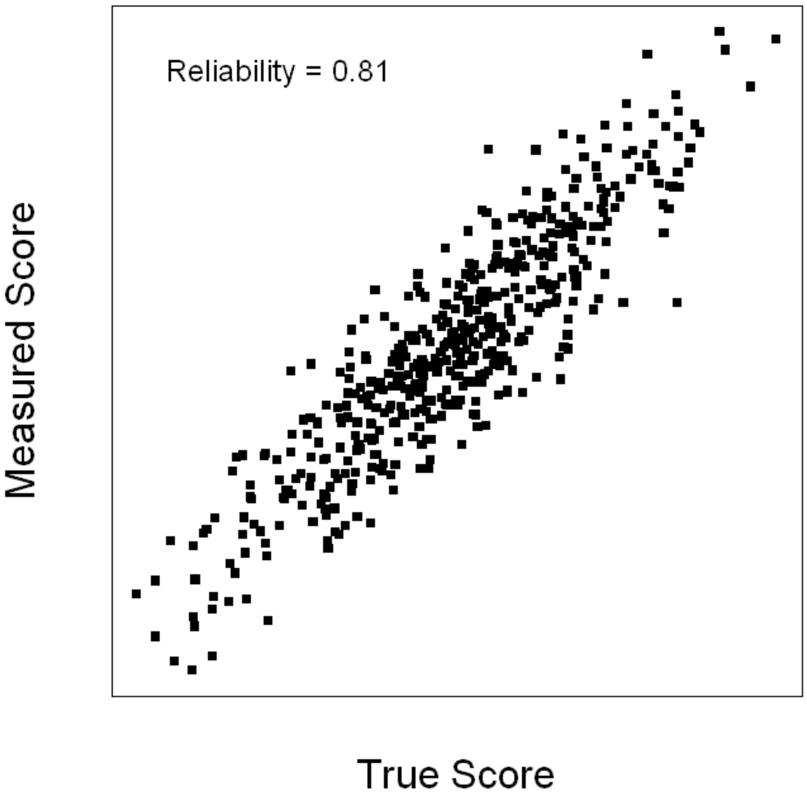
p^2 is the estimated squared correlation between the set of surgeon-specific estimates of composite performance measure values and their corresponding unknown true values (used as the measure of reliability in this study).

PrI = probability interval.

Based in part on these results, we selected a threshold of 100 cases over 3 years, as a minimum threshold for receiving a surgeon-specific composite score. This resulted in a reliability of 0.81 but reduced the number of surgeons eligible to receive a score from 2,286 to 1,976. A higher volume threshold would have yielded even higher reliability but at the cost of further reducing the number of surgeons 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 = 0.81. 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 $i = 1,2, \dots ,1976$ indicates the 1,976 surgeons 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 0.81.



____

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*)

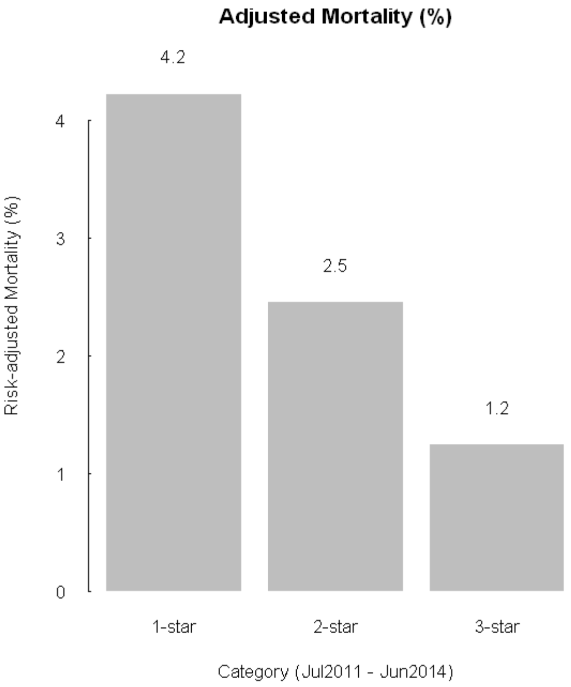
☐ **Empirical validity testing of the component measure score(s)**

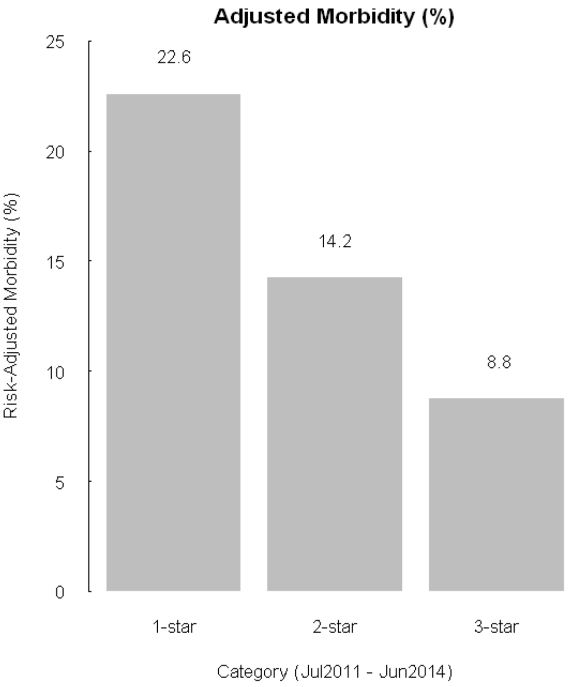
□ **Systematic assessment of face validity of component measure score(s) as an indicator** of quality or resource use (*i.e., is an accurate reflection of performance on quality or resource use and can distinguish good from poor performance*)

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: Surgeons were labeled as having higher-than-expected performance if the 98% credible interval surrounding a surgeon's composite score fell entirely above the overall STS average composite score. Surgeons were labeled as having lower-than-expected performance if the 98% credible interval surrounding a surgeon's composite score fell entirely below the overall STS average composite score. Surgeons 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.

2b1.3. What were the statistical results from validity testing? (*e.g., correlation; t-test*)
Compared to surgeons receiving 1 star, those with 3 stars had lower risk-adjusted mortality (1.2% vs. 4.2%) and lower risk-adjusted morbidity (8.8% vs. 22.6%) during July 2011 – June 2014. Thus, differences in performance were clinically meaningful as well as statistically significant. STS surgeons deemed better by the composite scores have (on average) higher performance during the same time window on each individual domain of the composite measure.





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 wide differences in risk-adjusted mortality and morbidity rates across categories of composite performance, and observed differences in morbidity and mortality rates correspond appropriately with the changes in performance categories. These results support the validity of the composite measure as a quality measure for cardiac surgery.

____

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*)

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*)

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*)

____

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 62 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?

Surgeon-specific risk-adjusted operative mortality and major complication rates were estimated using a bivariate random-effects logistic regression model [1]. To adjust for case mix, each patient's risk score for operative mortality and his or her corresponding risk score for major complications were first calculated, using existing and modified STS risk models as described below. The goal of calculating a risk score was to reduce the number of covariates in the hierarchical model by summarizing the predictive information from a large number of baseline covariates into a single number. Adjustment for each covariate individually in the hierarchical model would be theoretically preferable but is impractical due to the large number of records and covariates and the computationally intensive nature of Bayesian hierarchical model estimation.

To study the consequences of using an overall risk score for each patient instead of individual covariates in our models, sensitivity analyses were performed in which each surgeon's risk-adjusted mortality and complication rates were estimated in models that adjusted for 41 and 47 individual patient covariates, respectively. These estimates were compared with those derived from models adjusting for a single composite risk score. To make this analysis computationally manageable, model variables were estimated by maximum likelihood (i.e., empirical Bayes) instead of performing a fully Bayesian analysis. To further simplify this sensitivity analysis, mortality and complication rates were estimated in separate models, not simultaneously in a single model, and the cohort was restricted to isolated CABG. For each end point (operative mortality and major complications) we calculated each surgeon's risk-standardized rate of the end point using each model and compared the results.

After sensitivity analyses demonstrated the validity of this risk score approach, the operative mortality risk score (predicted risk of death) was then used as a covariate in the hierarchical model for operative mortality, and the major complication risk score (predicted risk of major morbidity) was used as a covariate in the hierarchical model for major complications. To reduce potential bias, the hierarchical model included both individual patient-level risk scores and the average value of these patient-level risk scores calculated separately for each surgeon [1].

For patients undergoing isolated CABG, isolated AVR, or AVR + CABG, risk scores were calculated according to the published STS 2008 mortality and major complications models for isolated CABG, isolated valve, or valve + CABG [2-4]. To ensure high calibration for the current study cohort, coefficients of each model were re-estimated using the current 3-year study sample and current end point definitions. Risk scores for patients undergoing a mitral operation without CABG were calculated using a modified version of the published STS 2008 mortality and major complications models for isolated valve procedures [3]. These modifications allowed inclusion of patients undergoing tricuspid repair, an increasingly common adjunct, urgent and emergency procedures, all arrhythmia ablation procedures for atrial fibrillation, atrial septal defect and patent foramen ovale closures, and active and treated endocarditis. Also included are more granular classifications and adjustment for the degree of tricuspid regurgitation (less than moderate, moderate, and severe).

Coefficients of the modified models were estimated using the current 3-year study cohort and end point definitions. Risk scores for patients undergoing a mitral operation with concomitant CABG were calculated using a similarly modified version of the published STS 2008 mortality and major complications models for valve + CABG operations [4], also with re-estimated coefficients.

References

1. Shahian DM, He X, Jacobs JP, Kurlansky PA, Badhwar V, Cleveland JC Jr, Fazzalari FL, Filardo G, Normand SL, Furnary AP, Magee MJ, Rankin JS, Welke KF, Han J, O'Brien SM. The Society of Thoracic Surgeons Composite Measure of Individual Surgeon Performance for Adult Cardiac Surgery: A Report of The Society of Thoracic Surgeons Quality Measurement Task Force. Ann Thorac Surg. 2015;100:1315-25.
2. Shahian DM, O'Brien SM, Filardo G, Ferraris VA, et al. The Society of Thoracic Surgeons 2008 cardiac surgery risk models: part 1--coronary artery bypass grafting surgery. Ann Thorac Surg. 2009 Jul;88(1 Suppl):S2-22.
3. O’Brien SM, Shahian DM, Filardo G, et al. The Society of Thoracic Surgeons 2008 cardiac surgery risk models: part 2—isolated valve surgery. Ann Thorac Surg 2009;88(1 Suppl):S23–42.
4. 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?

For isolated CABG, isolated AVR, and AVR+ CABG, risk adjustment is based on existing published risk models [1-3]. Methods and results for selecting covariates may be found in those publications.

For mitral valve repair and replacement (MVRR) and for MVRR + CABG, the published models were modified to account for the inclusion of patients undergoing tricuspid repair. Estimated odds ratios from these modified STS 2008 models are summarized in the tables below.

Odds ratios for the modified MVRR model

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	1.14 (1.08, 1.20)	<.0001	1.22 (1.09, 1.36)	0.0005
Race (v. others)	*	*	*	*

Black	1.25 (1.15, 1.35)	<.0001	*	*
Hispanic	1.23 (1.10, 1.39)	0.0003	*	*
CVD (v. no)	*	*	*	*
CVD with CVA	1.17 (1.08, 1.27)	0.0002	*	*
CVD without CVA	0.97 (0.88, 1.08)	0.6072	*	*
Number Diseased Vessels (3 v. 2, 2 v. 1/0)	1.06 (1.00, 1.12)	0.0681	*	*
Pre-op IABP or inotrope	2.20 (1.93, 2.52)	<.0001	1.35 (1.10, 1.66)	0.0040
Hypertension	1.13 (1.06, 1.20)	<.0001	1.11 (0.98, 1.27)	0.1115
Immunosuppressive treatment	1.26 (1.13, 1.41)	<.0001	1.60 (1.33, 1.93)	<.0001
Peripheral vascular disease	1.24 (1.14, 1.35)	<.0001	1.31 (1.13, 1.52)	0.0004
Aortic stenosis	1.12 (0.99, 1.27)	0.0714	*	*
MI <21 days	1.40 (1.19, 1.63)	<.0001	1.77 (1.41, 2.21)	<.0001
Shock	2.52 (2.08, 3.06)	<.0001	1.84 (1.38, 2.47)	<.0001
Number of previous operations (v. 0)	*	*	*	*
1 previous operation	1.35 (1.21, 1.50)	<.0001	1.69 (1.34, 2.15)	<.0001
2 or more previous operations	1.65 (1.39, 1.95)	<.0001	2.12 (1.53, 2.94)	<.0001
Urgent status (v. elective)	1.32 (1.23, 1.41)	<.0001	1.27 (1.12, 1.44)	0.0003
Active infections endocarditis	1.80 (1.63, 1.99)	<.0001	1.86 (1.54, 2.23)	<.0001
Treated infections endocarditis	1.07 (0.96, 1.19)	0.2471	0.96 (0.76, 1.22)	0.7427
Ejection fraction per 10-unit decrease	1.11 (1.07, 1.15)	<.0001	1.17 (1.09, 1.25)	<.0001
Creatinine per 1 unit increase	1.62 (1.53, 1.71)	<.0001	1.48 (1.36, 1.62)	<.0001
Body surface area, m^2	*	*	*	*
1.6 v. 2.0 in male	1.26 (1.11, 1.44)	0.0004	1.64 (1.29, 2.09)	<.0001
1.8 v. 2.0 in male	1.05 (1.00, 1.10)	0.0418	1.19 (1.08, 1.30)	0.0002
2.2 v. 2.0 in male	1.09 (1.05, 1.13)	<.0001	0.99 (0.91, 1.06)	0.6966
1.6 v. 1.8 in female	1.07 (1.03, 1.12)	0.0017	1.25 (1.16, 1.35)	<.0001
2.0 v. 1.8 in female	1.04 (1.01, 1.08)	0.0074	0.98 (0.92, 1.04)	0.4325
2.2 v. 1.8 in female	1.21 (1.12, 1.31)	<.0001	1.17 (1.00, 1.37)	0.0500
Time trend (half year increase)	0.96 (0.95, 0.98)	<.0001	1.01 (0.98, 1.04)	0.6571
Left main disease	*	*	1.11 (0.81, 1.53)	0.5091
Unstable angina (no MI < 8days)	*	*	1.21 (0.94, 1.55)	0.1420
Mitral stenosis	*	*	1.05 (0.91, 1.20)	0.5060
Moderate tricuspid insufficiency (v. no-mild)	1.13 (1.05, 1.20)	0.0003	1.17 (1.02, 1.33)	0.0243
Severe tricuspid insufficiency (v. no-mild)	1.23 (1.12, 1.35)	<.0001	1.46 (1.22, 1.75)	<.0001

Mitral valve repair (v. replacement)	0.56 (0.50, 0.62)	<.0001	0.40 (0.31, 0.53)	<.0001
Tricuspid valve repair (v. none)	1.36 (1.24, 1.49)	<.0001	0.99 (0.84, 1.18)	0.9474
Effects that interacts with procedure groups and were modeled separately for MV replacement and MV repairs in MV replacements:	*	*	*	*
Age	*	*	*	*
60 v. 50 (no reoperations, non-emergent)	1.22 (1.17, 1.27)	<.0001	1.53 (1.40, 1.67)	<.0001
70 v. 50 (no reoperations, non-emergent)	1.49 (1.37, 1.62)	<.0001	2.34 (1.95, 2.81)	<.0001
80 v. 50 (no reoperations, non-emergent)	1.80 (1.61, 2.01)	<.0001	3.69 (2.95, 4.63)	<.0001
Congestive heart failure (v. no)	*	*	*	*
CHF not NYHA IV	1.08 (1.00, 1.17)	0.0517	1.20 (1.06, 1.37)	0.0043
CHF NYHA IV	1.53 (1.38, 1.69)	<.0001	1.66 (1.40, 1.96)	<.0001
Diabetes (v. no)	*	*	*	*
Insulin diabetes	1.41 (1.27, 1.57)	<.0001	1.48 (1.25, 1.75)	<.0001
Non-insulin diabetes	1.10 (1.02, 1.20)	0.0174	1.14 (1.00, 1.31)	0.0587
Chronic lung disease (severe v moderate, or moderate v none-mild)	1.15 (1.11, 1.18)	<.0001	1.20 (1.13, 1.28)	<.0001
Dialysis v. no dialysis & creatinine = 1.0	1.97 (1.75, 2.22)	<.0001	2.59 (2.12, 3.15)	<.0001
Female (at BSA=1.8) v. male (at BSA=2.0)	1.17 (1.11, 1.25)	<.0001	1.32 (1.13, 1.53)	0.0004
Status (v. elective)	*	*	*	*
Emergent - no resuscitation	3.30 (2.55, 4.27)	<.0001	2.38 (1.61, 3.49)	<.0001
Emergent+resuscitation/emergent salvage	2.83 (1.63, 4.89)	0.0002	5.91 (3.18, 10.98)	<.0001
In MV repairs:	*	*	*	*
Age	*	*	*	*
60 v. 50 (no reoperations, non-emergent)	1.27 (1.21, 1.32)	<.0001	1.76 (1.57, 1.97)	<.0001
70 v. 50 (no reoperations, non-emergent)	1.60 (1.47, 1.75)	<.0001	3.09 (2.46, 3.88)	<.0001
80 v. 50 (no reoperations, non-emergent)	2.00 (1.78, 2.25)	<.0001	5.61 (4.19, 7.50)	<.0001
Congestive heart failure (v. no)	*	*	*	*
CHF not NYHA IV	1.17 (1.07, 1.28)	0.0007	1.20 (1.06, 1.37)	0.0043
CHF NYHA IV	1.55 (1.36, 1.76)	<.0001	1.66 (1.40, 1.96)	<.0001
Diabetes (v. no)	*	*	*	*
Non-insulin diabetes	1.10 (0.99, 1.21)	0.0690	1.14 (1.00, 1.31)	0.0587
Insulin diabetes	1.44 (1.24, 1.66)	<.0001	1.48 (1.25, 1.75)	<.0001

Chronic lung disease (severe v. moderate, or moderate v. none-mild)	1.15 (1.11, 1.18)	<.0001	1.26 (1.16, 1.38)	<.0001
Dialysis v. no dialysis & creatinine = 1.0	1.97 (1.75, 2.22)	<.0001	3.68 (2.44, 5.54)	<.0001
Female (at BSA=1.8) v. male (at BSA=2.0)	1.17 (1.11, 1.25)	<.0001	1.14 (0.93, 1.40)	0.2108
Status (v. elective)	*	*	*	*
Emergent - no resuscitation	3.30 (2.55, 4.27)	<.0001	3.83 (1.87, 7.86)	0.0003
Emergent+resuscitation/Emergent Salvage	2.83 (1.63, 4.89)	0.0002	2.47 (0.10, 59.60)	0.5785

*cell intentionally left blank

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.

Odds ratios for the modified MVRR + CABG model

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)	*	*	*	*
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)	*	*	*	*

Effect	Morbidity: OR (95% CI)	Morbidity: P-value	Mortality: OR (95% CI)	Mortality: P-value
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
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	*	*	*	*

Effect	Morbidity: OR (95% CI)	Morbidity: P-value	Mortality: OR (95% CI)	Mortality: P-value
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
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

*cell intentionally left blank

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.

References

1. Shahian DM, O'Brien SM, Filardo G, Ferraris VA, et al. The Society of Thoracic Surgeons 2008 cardiac surgery risk models: part 1--coronary artery bypass grafting surgery. Ann Thorac Surg. 2009 Jul;88(1 Suppl):S2-22.
2. O’Brien SM, Shahian DM, Filardo G, et al. The Society of Thoracic Surgeons 2008 cardiac surgery risk models: part 2—isolated valve surgery. Ann Thorac Surg 2009;88(1 Suppl):S23–42.
3. 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.

2b3.4b. 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*)

For isolated CABG, isolated AVR, and AVR+ CABG, risk adjustment is based on existing published risk models [1-3]. Model performance metrics may be found in those publications.

For MVRR, modifications to the existing STS models were assessed using data from 62,118 patients undergoing MVRR during July 2011 – June 2014.

For MVRR + CABG, modifications to the existing STS models were assessed using data from 26,355 patients undergoing MVRR + CABG during July 2011 – June 2014.

Discrimination

Discrimination results are presented for the modified MVRR and MVRR + CABG models. 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 [4].

Calibration

Calibration results are presented for the modified MVRR and MVRR + CABG models. 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.

References

1. Shahian DM, O'Brien SM, Filardo G, Ferraris VA, et al. The Society of Thoracic Surgeons 2008 cardiac surgery risk models: part 1--coronary artery bypass grafting surgery. Ann Thorac Surg. 2009 Jul;88(1 Suppl):S2-22.
2. O'Brien SM, Shahian DM, Filardo G, et al. The Society of Thoracic Surgeons 2008 cardiac surgery risk models: part 2—isolated valve surgery. Ann Thorac Surg 2009;88(1 Suppl):S23–42.
3. 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.

4. 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*)**:**

Modified MVRR model

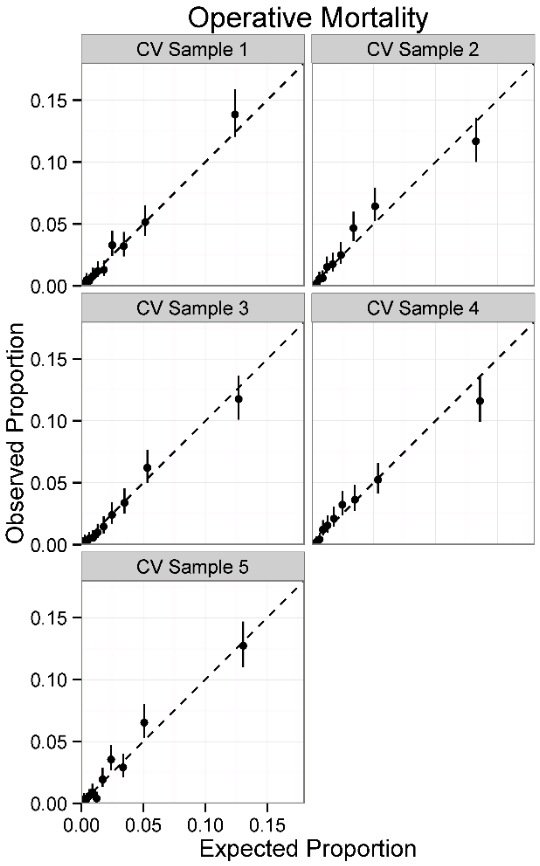
The bootstrap-adjusted estimated C-statistic was 0.746 for the morbidity model and 0.807 for the mortality model. These numbers were comparable to the STS 2008 models when evaluated using the same sample (0.745 and 0.807 for morbidity and mortality endpoints, respectively.)

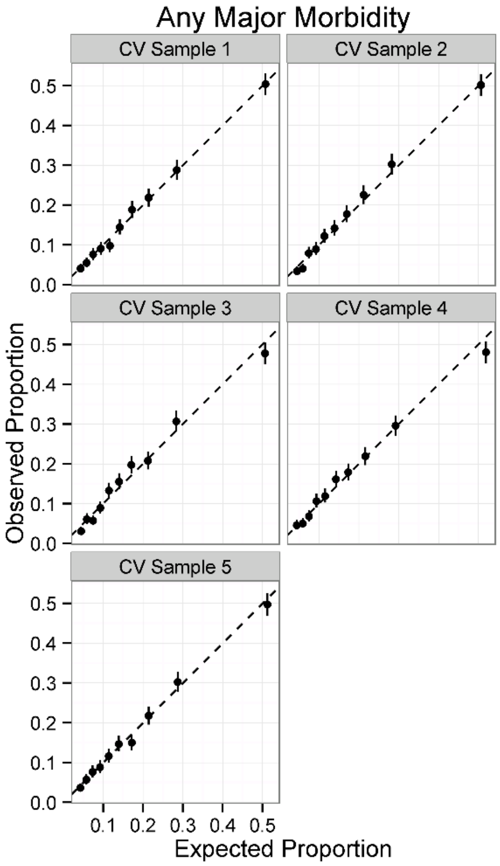
Modified MVRR + CABG model

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.)

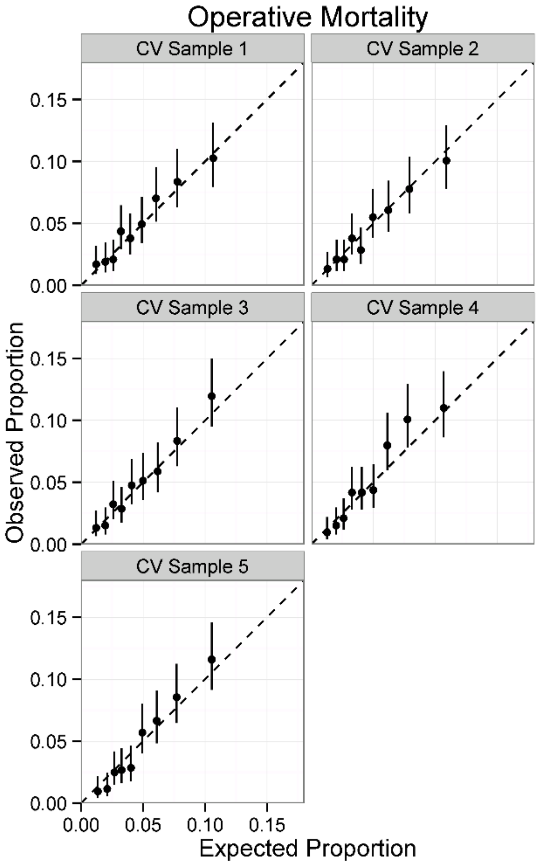
2b3.7. Statistical Risk Model Calibration Statistics (*e.g., Hosmer-Lemeshow statistic*):
N/A. The Hosmer-Lemeshow statistic was not calculated.

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





Modified MVRR + CABG model



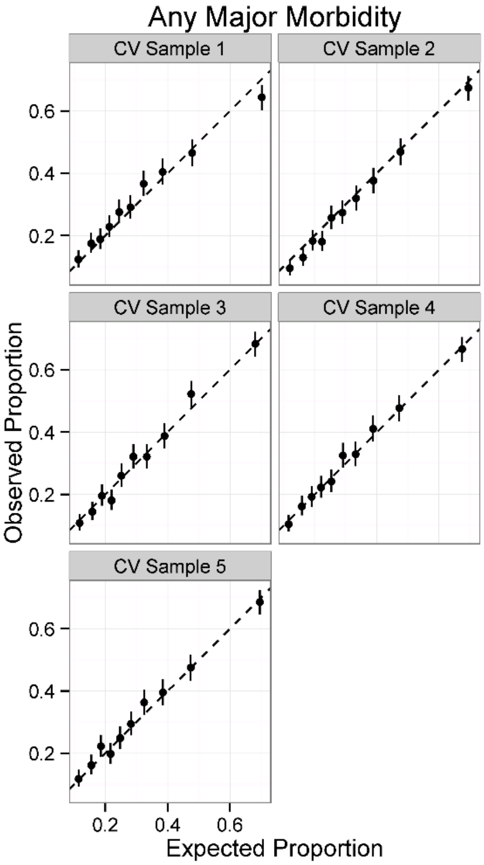


Figure. Plots of observed versus expected in cross validation samples, operative mortality

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 surgeon's composite measure estimate is indicated by calculating 98% Bayesian credible intervals (CI's) which are similar to conventional confidence intervals. Point estimates and CI's for an individual STS surgeon 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, 2 and 3 stars. An STS surgeon receives 2 stars if the Bayesian credible interval surrounding his/her composite score overlaps the overall STS average. This rating implies that the STS surgeon's performance was not statistically different from the overall STS national average. If the Bayesian CI falls entirely above the STS national average, the surgeon receives 3 stars (higher-than-expected performance). If the Bayesian CI falls entirely below the STS national average, the surgeon 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 surgeons with at least 100 cases over 3 years, around 71% of surgeons received 2 stars, and the remaining surgeons received either 1 or 3 stars.

Performance categories July 2011 – June 2014

Category	All Surgeons: Number of Surgeons, %	Surgeons N≥ 100: Number of Surgeons, %
1-star	207, 9.1%	189, 9.6%
2-star	1701, 74.4%	1413, 71.5%
3-star	378, 16.5%	374, 18.9%

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 surgeons 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.*

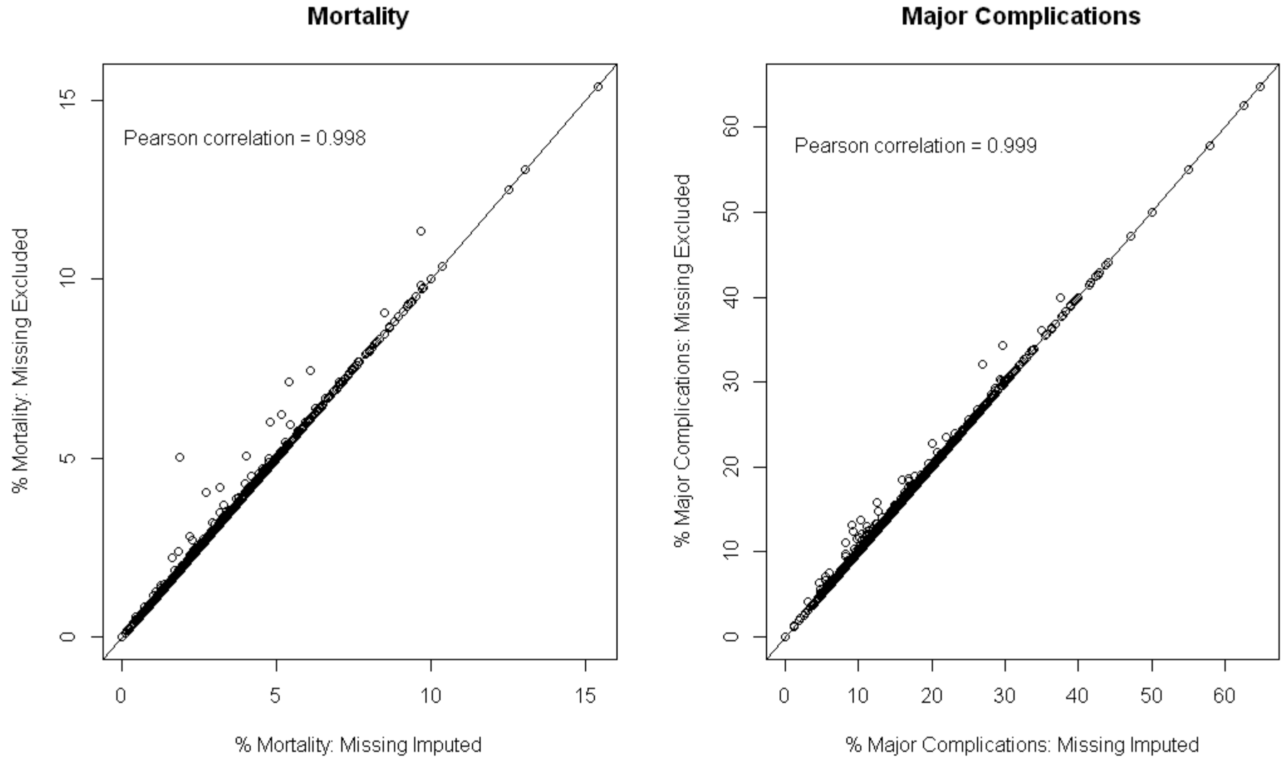
2b6.1. Describe the method of testing conducted to identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and non-responders) and how the specified handling of missing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

Missing data for risk model covariates were extremely rare: All model predictors had <5% missing and the majority had <1% missing. Missing data occurred in 0.4% of records for operative mortality and 0.3% 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 0.4% for operative mortality and 0.3% for major complications. The median surgeon-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 surgeons with >10% missing data was 1.0% for mortality and 1.0% for major complications. As a sensitivity analysis, we re-calculated surgeon-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 surgeon-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 surgeons.

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
0.73	0.92

The Pearson correlations were 0.73 for mortality versus overall composite measure and 0.92 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 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)*

The overall composite score was calculated as a weighted sum of (1 minus risk-adjusted mortality rate) and (1 minus risk-adjusted major morbidity rate). Mortality and morbidity rates were weighted inversely by their respective standard deviations across surgeons. This procedure is equivalent to first rescaling mortality and morbidity rates by their respective standard deviation across surgeons and then assigning equal weighting to the rescaled mortality rate and rescaled morbidity rate. Standard deviations derived from the data were used to define the final composite measure as 0.81 × (1 minus risk-standardized mortality rate) + 0.19 × (1 minus risk-standardized complication rate).

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 0.81 and 0.19, respectively. An implication of this weighting is that a 1 percentage point change in a surgeon's risk-adjusted mortality rate has the same impact as a 4.3 percentage point change in the surgeon'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.

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.

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	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

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

Confidential, surgeon-level performance results for this composite measure were first distributed to consenting surgeons participating in the STS Adult Cardiac Surgery Database in January 2020.
This measure is not yet publicly reported; see response under 4a1.2.

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?*)

Although this composite was developed in 2014 and first endorsed by NQF in 2017, concerns related to the confidentiality and format of the surgeon-level results delayed the initial annual distribution of performance reports until Jan., 2020. STS requires our measured entities, whether individual surgeons or surgeon groups, to have multiple opportunities to receive and provide feedback on performance reports before adding new measures to our public reporting program. Public reporting for the individual surgeon composite will therefore not be under consideration until 2021 or 2022.

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.*)

Please see response under 4a1.2

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.

Please see 1b.2 (Importance tab) for description of the identification of 2,098 surgeons who met the completeness and minimum procedure thresholds, 1,841 of whom performed at least 100 eligible cases within the three-year measurement period. Of this subset of surgeons, approximately 400 opted in for receipt of their confidential, surgeon-level performance results in January 2020.

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.

The annual performance reports include separate mortality and morbidity domain scores and an overall composite score. The surgeon´s score is illustrated graphically in relation to the 10th, 50th and 90th percentiles of the distribution across all surgeons who were eligible for inclusion in the analysis for the specified three-year period, and is also accompanied by the 98% Bayesian credible interval. A detailed report overview, providing explanations of statistical calculations, endpoints, and report interpretation, is included in the report.

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. and Canada who comprise the STS Adult Cardiac Surgery Database and Quality Measurement Task Forces meet periodically to discuss the surgeon-level and participant reports and to consider potential enhancements to the ACSD. Additions/clarifications to the data collection form and to the content/format of the individual surgeon reports and participant reports are discussed and implemented as appropriate.

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

N/A (performance results for this measure are shared with consenting surgeons only)

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.

Performance data for the Individual Surgeon Composite (3030) were first distributed to consenting surgeons in January 2020; overall performance trends for this measure are therefore not yet available. As noted elsewhere in these submission materials, measure 3030 aggregates individual surgeon performance on five surgical procedures (isolated CABG, isolated AVR, AVR+CABG, MVRR, MVRR+CABG) and provides each surgeon with mortality and morbidity domain scores and an overall composite score and star rating, based on their own case mix. Therefore, in the absence of multi-year performance trends for measure 3030, we are providing (below) the star rating trends for the five procedures aggregated within it.

The data demonstrate that the general trend since the introduction of each measure has been a decrease in the percentage of surgical programs with 1-star and 3-star ratings and a corresponding increase in 2-star programs. This trend is consistent with the performance improvement goals of the STS star rating program, which seek to reduce variation in performance and to drive all participants in the STS Adult Cardiac Surgery Database toward the 2-star (or "as expected") category.

(If table below does not display clearly, please see version in Appendix.)

	Stars	2019	2018	2017	2016	2015	2014	2013	2012	2011	2010	
CABG	*	3.77	4.37	4.55	5.29	5.82	4.59	9.19	9	9.6	11	
	**	88.57	88.27	89.21	84.65	84.4	86.64	75.86	76	76.5	75.5	
	***	7.66	7.36	6.24	10	9.74	8.77	14.95	15	14	13.5	
AVR	*	1.67	1.96	2.62	2.17	3.11	4.22	3.35	3	3.5	N/A	
	**	92.26	92.84	92.70	90.3	88.75	87.89	88.98	91	90.6	N/A	
	***	6.07	5.20	4.68	7.53	8.15	7.89	7.67	6	5.9	N/A	
AVR + CABG		*	1.84	2.16	2.73	2.06	2.49	2.51	3.14	N/A	N/A	N/A
	**	93.5	93.03	92.76	92.26	90.72	90.42	90.7	N/A	N/A	N/A	
	***	4.66	4.81	4.51	5.68	6.79	7.07	6.17	N/A	N/A	N/A	
MVRR	*	1.85	2.41	3.64	N/A	N/A	N/A	N/A	N/A	N/A	N/A	
	**	91.81	87.06	85.65	N/A	N/A	N/A	N/A	N/A	N/A	N/A	

	***	6.34	10.53	10.71	N/A	N/A	N/A	N/A	N/A	N/A	N/A	
MVRR + CABG		*	2.55	2.08	2.74	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	**	88.0	89.97	91.78	N/A	N/A	N/A	N/A	N/A	N/A	N/A	
	***	9.45	7.96	5.48	N/A	N/A	N/A	N/A	N/A	N/A	N/A	

Performance data for the Individual Surgeon Composite (3030) were first distributed to consenting surgeons in January 2020; overall performance trends for this measure are therefore not yet available. As noted elsewhere in these submission materials, measure 3030 aggregates individual surgeon performance on five surgical procedures (isolated CABG, isolated AVR, AVR+CABG, MVRR, MVRR+CABG) and provides each surgeon with mortality and morbidity domain scores and an overall composite score and star rating, based on their own case mix. Therefore, in the absence of multi-year performance trends for measure 3030, we are providing (below) the star rating trends for the five procedures aggregated within it.

The data demonstrate that the general trend since the introduction of each measure has been a decrease in the percentage of surgical programs with 1-star and 3-star ratings and a corresponding increase in 2-star programs. This trend is consistent with the performance improvement goals of the STS star rating program, which seek to reduce variation in performance and to drive all participants in the STS Adult Cardiac Surgery Database toward the 2-star (or "as expected") category.

(If table below does not display clearly, please see version in Appendix.)

	Stars	2019	2018	2017	2016	2015	2014	2013	2012	2011	2010	
CABG	*	3.77	4.37	4.55	5.29	5.82	4.59	9.19	9	9.6	11	
	**	88.57	88.27	89.21	84.65	84.4	86.64	75.86	76	76.5	75.5	
	***	7.66	7.36	6.24	10	9.74	8.77	14.95	15	14	13.5	
AVR	*	1.67	1.96	2.62	2.17	3.11	4.22	3.35	3	3.5	N/A	
	**	92.26	92.84	92.70	90.3	88.75	87.89	88.98	91	90.6	N/A	
	***	6.07	5.20	4.68	7.53	8.15	7.89	7.67	6	5.9	N/A	
AVR + CABG		*	1.84	2.16	2.73	2.06	2.49	2.51	3.14	N/A	N/A	N/A
	**	93.5	93.03	92.76	92.26	90.72	90.42	90.7	N/A	N/A	N/A	
	***	4.66	4.81	4.51	5.68	6.79	7.07	6.17	N/A	N/A	N/A	
MVRR	*	1.85	2.41	3.64	N/A	N/A	N/A	N/A	N/A	N/A	N/A	
	**	91.81	87.06	85.65	N/A	N/A	N/A	N/A	N/A	N/A	N/A	
	***	6.34	10.53	10.71	N/A	N/A	N/A	N/A	N/A	N/A	N/A	
MVRR + CABG		*	2.55	2.08	2.74	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	**	88.0	89.97	91.78	N/A	N/A	N/A	N/A	N/A	N/A	N/A	
	***	9.45	7.96	5.48	N/A	N/A	N/A	N/A	N/A	N/A	N/A	

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

3032 - Mitral Valve Repair/Replacement + CABG 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_Surgeon_Composite_Appendix_-_S.4-11-14-15_1b.2-_1b.4-_10262020-637408683591885628.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: 2015

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