NATIONAL QUALITY FORUM

MEASURE WORKSHEET

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

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

Brief Measure Information

NQF #: 0696

Corresponding Measures:

De.2. Measure Title: STS CABG Composite Score

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

De.3. Brief Description of Measure: The STS CABG Composite Score comprises four domains consisting of 11 individually NQF-endorsed cardiac surgery measures:

Domain 1) Absence of Operative Mortality – Proportion of patients (risk-adjusted) who do not experience operative mortality. Operative mortality is defined as death during the same hospitalization as surgery or after discharge but within 30 days of the procedure;

Domain 2) Absence of Major Morbidity – Proportion of patients (risk-adjusted) who do not experience any major morbidity. Major morbidity is defined as having at least one of the following adverse outcomes: 1. reoperations for any cardiac reason, 2. renal failure, 3. deep sternal wound infection, 4. prolonged ventilation/intubation, 5. cerebrovascular accident/permanent stroke;

Domain 3) Use of Internal Mammary Artery (IMA) – Proportion of first-time CABG patients who receive at least one IMA graft;

Domain 4) Use of All Evidence-based Perioperative Medications – Proportion of patients who receive all required perioperative medications for which they are eligible. The required perioperative medications are: 1. preoperative beta blockade therapy, 2. discharge anti-platelet medication, 3. discharge beta blockade therapy, and 4. discharge anti-lipid medication.

All measures are based on audited clinical data collected in a prospective registry. Participants receive a score for each of the domains, plus an overall composite score. The overall composite score is created by "rolling up" the domain scores into a single number. In addition to receiving a numeric score, participants are assigned to rating categories designated by one star (below average performance), two stars (average performance), or three stars (above average performance). For consenting participants, scores and star ratings are publicly reported on the STS website.

1b.1. Developer Rationale: N/A

S.4. Numerator Statement: Please see Appendix

S.6. Denominator Statement: Please see Appendix

S.8. Denominator Exclusions: Please see Appendix

De.1. Measure Type: Composite

S.17. Data Source: Registry Data

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

Preliminary Analysis: Maintenance of Endorsement

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

Criteria 1: Importance to Measure and Report

1a. Evidence

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

1a. Evidence. The evidence requirements for a health outcome measure include providing empirical data that demonstrate a relationship between the outcome and at least one healthcare structure, process, intervention, or service; 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 is comprised of 11 NQF-endorsed STS measures that are included in 4 domains representing outcomes of absence of mortality and absence of any major morbidity (defined as one of five adverse outcomes) as well as process measures for use of internal mammary artery (IMA) graft and use of all evidence-based medications for which they are eligible (preoperative beta blockade, discharge beta blockade, discharge anti-platelet, discharge anti-lipid).
- The developer stated that the composite measure provides a more comprehensive measure of overall performance/quality than does a single measure of mortality.
- The approach to development of the model, including decision logic and results of testing (with STS registry data) used to combine the data from each of the measures into respective domains and then to combine domain scores into a single composite score, was presented and described in detail in a paper that addresses composite measure scoring and provider ratings.
- References that address mortality and the major morbidity across the component measures were provided. A study using data from over 500,000 CABG procedures from the STS National Adult Cardiac Surgery Database in the period 1997 – 1999 reported 30 day mortality rates (3.05%) and major complications (stroke, renal failure, reoperation, prolonged ventilation, sternal infection) rates (13.40%). The authors concluded that, while the correlation was slight, when used in combination, mortality and morbidity indicators may provide information that would help surgery teams evaluate the quality of their care and allow them to focus on areas of improvement. Other studies have reported that while mortality rates are coming down, quality of care may be better understood, and improved, by also looking at morbidity.

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**

Questions for the Committee:

- *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?*

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

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

Maintenance measures – increased emphasis on gap and variation

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

- Distribution of STS isolated CABG measure from two consecutive time periods, July 2015 – June 2016 and July 2016 – June 2017, the mean composite score for 4 "harvests" during those periods are 0.967 (latest), 0.967, 0.967, and 0.966. Composite score distribution within each harvest ranged from 0.923 to 0.987 (latest), 0.891 to 0.986 (Spring 2014), 0.900 to 0.987 (Fall 2013), and 0.899 to 0.986 (Spring 2013).

Distribution of STS isolated CABG composite measure in the latest four STS harvests for which the measure was reported (US Geographic Region)

	Latest	**Spring 2017**	**Fall 2016**	**Spring 2016**
# of Participants	945	1,006	882	1,026
# of Operations	145,815	150,882	129,972	149,917
Mean	0.967	0.967	0.967	0.966
STD	0.00972	0.0109	0.0102	0.0104
IQR	0.0123	0.0142	0.0131	0.0134
Min (0^{th})	0.919	0.923	0.917	0.912
10^{th}	0.954	0.952	0.954	0.953
20^{th}	0.959	0.958	0.960	0.958
30^{th}	0.962	0.962	0.964	0.962
40^{th}	0.965	0.965	0.966	0.965
50^{th}	0.968	0.968	0.969	0.968
60^{th}	0.970	0.971	0.971	0.970
70^{th}	0.972	0.973	0.974	0.973
80^{th}	0.975	0.976	0.976	0.975
90^{th}	0.978	0.980	0.978	0.978
Max (100^{th})	0.985	0.989	0.986	0.986

Disparities

- Disparities data is presented for participant (hospital, hospital group, surgeon group) rather than patient thus disparities are provided for distribution of results by region across risk-adjusted odds ratios for race and sex for the 4 domains: mortality, morbidity, IMA use, and perioperative medication use.

Risk-adjusted odds ratios associated with sex and race at the individual domain level (Fall 2016 harvest, data from June 2015 – June 2016)

	Risk-adjusted odds ratio of mortality	**Risk-adjusted odds ratio of morbidity**
Female (at BSA = 1.8) v male (at BSA = 2.0)	1.59 (95% CI: 1.45-1.74)	1.30 (95% CI: 1.24-1.36)

Black v white (including patients with race other than white, black, Asian)	1.17 (95% CI: 1.03-1.32)	1.27 (95% CI: 1.18-1.36)
Asian v white (including patients with race other than white, black, Asian)	0.97 (95% CI: 0.80-1.19)	1.16 (95% CI: 1.04-1.30)

Observed proportions of IMA use and perioperative medications (Fall 2016 harvest, data from June 2015 – June 2016)

	Observed proportions of IMA use	**Observed proportions of perioperative medications**
Female v male	98.5% v 99.2%	92.6% v 92.5%
Black v non-black	98.7% v 99.1%	93.2% v 92.5%

Questions for the Committee:

- Is there a gap in care that warrants a national performance measure?
- Does the approach to disparities analysis assist in the understanding of the performance of the measure as specified?

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

1c. Composite – Quality Construct and Rationale

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

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

- This composite measure is comprised of 11 measures that are grouped into four domains (mortality, morbidity, IMA use, perioperative medication use) with an individual score for each domain and a score for the composite resulting from rolling up the 4 domain scores.
    - The morbidity domain, which consists of 5 measures, is scored "any or none" meaning that occurrence of any one of the 5 adverse events determines the score.
    - The perioperative medication domain is scored "all-or-none" meaning all evidence-based perioperative medications included in the domain are received by each patient.
    - The remaining two domains are each comprised of a single measure (IMA use and mortality) and each is scored as a proportion.
    - The domains are combined into a composite based on weights reflecting their importance: 81% of total weight applied to mortality, 10% to morbidity, 7% to IMA, and 3% to medications.
- The developers provide the following rationale for the composite: "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. This composite provides a more comprehensive measure of overall quality."
- The aggregation method for the composite is at the patient level. Overall composite performance is calculated as a weighted average of the domain-specific estimates and reflected at the participant level.

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 clearly articulated in the submission and logical?

Preliminary rating for composite quality construct and rationale:

☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

RATIONALE:

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."

- I am not aware of any new evidence that would change the evidence base for this measure. The evidence base for mortality and major morbidity measures are more or less self-evident. There is also no new data that would suggest that IMA implantation is not more efficacious than saphenous vein, nor that the appropriate perioperative medication regimen is not associated with a reduction in the risk of perioperative thromboembolic, cardiac & neurologic events.
- Given results of the ISCHEMIA trial on overuse of CABG in specific patients (JAMA Cardiol. 2019;4(3):273-286. doi:10.1001/jamacardio.2019.0014), would the measure benefit from inclusion of domains around appropriateness of CABG?
- Agree
- The composite is composed of NQF endorsed outcome and process measures for which underlying evidence is adequate
- The evidence if directly related to the measured outcome. I am unaware of any new data other than what the developer sites.
- I agree that the evidence basis for this composite measure has not changed.
- pass
- It is feasible, readily assessible without undue burden, and addresses patient's goal of improved quality of life with decrease of postop complications both short-term and long-term

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?

- Yes, although the performance distribution among centers appeared to be very "tight" (median performance = 96.8% IQR +/- 1.2%), there was both meaningful opportunity for improvement overall and significant variability across STS units across at least 3 of the four measures (mortality, morbidity and medications). IMA was used in about 99% of all patients; it's unclear to me how important this is in the overall assessment given that there appears to be limited variation among centers and overall it doesn't seem to be a compelling deficiency in STS CABG care. This is in contrast to morbidity and medication compliance where the greatest relative potential for improvement appears to be concentrated. The data on disparities is also compelling across the individual domains, with increased risk for M&M demonstrated for female sex and AA race, among

others. It is not clear to me what is meant by "we do not provide data stratified by patient characteristics, instead we provide results stratified by participant characteristics". It would seem that each hospital's overall composite score for their entire patient population would adjust for these factors, but also they would get a report stratified by high-risk characteristics as well- especially at hospitals that may manage a relatively high % of such high-risk patients (eg, inner-city hospitals).

- n/a
- Agree
- The range of performance is between ~92% and 98%. I'm concerned that the range of performance is too high and narrow.
- A moderate gap is highlighted.
- There continues to be an opportunity for improvement, little shift over the last 4 periods. Appropriate approach to disparities analysis demonstrating gap between sex and race in morbidity and mortality.
- moderate
- Population subgroups provided but limited. Gaps documented. Not a true representation of National population

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?

- The weighting scheme seems to be very well documented and supported by published consensus statements. Furthermore, the content validity assessment demonstrated that there were significant (and clinically meaningful differences) in all 4 domains (including IMA %'s surprisingly) between hospitals with 1 and 3 stars. However, despite the content validity, the rules were developed from an "expert panel". I would be curious how this panel came to this relative weighting scheme to assign the final composite score (and star ratings). How were these weights developed? Cost associated with different events? Excess hospital days? Based on the relative weighting, a 1% change in mortality would have the same impact on the overall composite score as a 11.6% change in morbidity. Were there any patients/families weighing in on the relative value of these different domains? I'm sure many patients would rather risk mortality vs. a severe debilitating stroke or massive MI. With these considerations, I'm wondering whether this composite score is meant to replace the individual four domain scores, or simply be used as a summary assessment? It should be the latter so that the more granular assessments are still available to STS units to decide upon where to prioritize their QI efforts and to patients so they can weigh what's important to them in choosing a hospital.
- Yes
- Abstain
- The rationale for the measure was that the mortality measure was topping out, thus motivating the inclusion of domains for which performance is more variable. The rationale and methods for the weighing are given. Perhaps the weighing should be reconfigured to produce more variability?
- The composite score is clearly described and appropriately weighted with distinct rationale for each individual component.
- Quality construct and rationale clear and logical as well as the method for aggregation and weighting.
- moderate

- Four domains clearly defined and weighted

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**

Methods Panel Review (Combined)

Methods Panel Evaluation Summary:

This measure was reviewed by the Scientific Methods Panel (SMP) Subgroup and discussed during the full SMP October 2019 In-Person Meeting. The Subgroup passed the measure on reliability. The Subgroup was initially unable to reach consensus regarding validity. During the in-person meeting, the full Panel discussed validity and then the Subgroup voted again, passing the measure on validity. A summary of the measure and the Panel discussion is provided below.

Reliability 0-H; 6-M; 1-L; 0-I → Measure passes with MODERATE rating

- Performance level signal to noise (SNR) with Bayesian approach to calculation true probability for the reliability testing. SNR= 0.68 with considerable spread.

Validity (final vote) 2-H; 4-M; 0-L; 0-I → Measure passes with MODERATE rating

- Measure gap: recent cohort of 1,024 practices (2014) showed performance of .97 (SD= .00092) across 143K procedures (presumed weighted and adjusted).
- Pearson's correlation between two time periods July 2013-2014 vs. July 2012-2013 was 0.64; Spearman's 0.63.
- Face validity results suggested, but not described in great detail.
- Meaningful differences somewhat evident, even as about 80% perform at average levels (see table below).

Table X. Star ratings in the last four harvests

Star Rating	07/2013-06/2014	01/2013-12/2013	07/2012-06/2013	01/2012-12/2012
1	60, 5.9%	98, 9.6%	97, 9.6%	91, 9.0%
2	864, 84.4%	770, 75.7%	782, 77.3%	770, 76.5%
3	100, 9.8%	149, 14.7%	132, 13.1%	146, 14.5%

- Correlations between star rating and domain score also presented, but arguably a bit circular in logic.
    - **Ratings for Composite Construction:** H-3; M-2; L-1; I-1 → Passes with HIGH Rating
        - 11 measures that lie beneath all are NQF endorsed
        - Compilation method that all must be achieved, for numerator of composite to be fulfilled
        - Weighted construction: see note under specifications
- Missing data: rare 1/1,000 observations, and only 13 of centers had more the 5% missing data, and results did not change.
- Key concerns in the SMP's initial analysis regarding the measure include the ability to detect meaningful differences in performance, questions about the appearance of increased scores for participants with more exclusions, questions about the age of the risk model, and that no external standard was used to demonstrate validity.
- In response to the concerns raised, the developer provided additional information regarding the lack of external metrics for validity testing, an updated risk model calibration, and updated validity data. The committee discussed the updated information and the NQF criteria. On re-vote, the measure passed validity with a moderate rating.

Questions for the Committee regarding reliability:

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

Questions for the Committee regarding validity:

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

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 Scientific Methods Panel 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 specific concerns— the data elements and risk-adjustment methodology has been previously vetted by the group (I believe) when the individual domains were assessed.
- n/a
- Abstain
- Specifications are clear
- All the data elements can reliably be abstracted and the measure is already being implemented.
- No concerns that measure can be consistently implemented. All elements and logic clearly defined.
- moderate
- No concerns at this time

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

- The reliability testing appears to be both comprehensive and very sophisticated, but admittedly above my pay grade in understanding. I will defer to those with sufficient expertise to comment further.
- n/a
- No
- As noted by the SMP, the overall performance has increased and the range has narrowed since the reliability analysis was conducted (2014). Therefore, the reliability analysis based on older data may not reflect current reality.
- No concerns.
- No concerns about reliability.
- no issues : moderate
- No

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

- As mentioned previously, the content validity assessment demonstrated that there were significant (and clinically meaningful differences) in all 4 domains (including IMA %'s surprisingly) between hospitals with 1 and 3 star ratings based on the composite measure. Furthermore, the measure

demonstrated strong empirical and face validity through stable/consistent hospital-level performance scores over time and utilization by Consumer Reports, respectively.

- Have the star ratings been tested with patients considering CABG and did they align with consumer understandings of what institute a 1 star vs 3 star facility?
- No
- As noted by the SMP, empirical testing analyses (required for maintenance) pertains to validity, rather I think it is testing lagged test-retest reliability (stability).
- No concerns.
- No concerns about validity.
- no issues: moderate
- No

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

- No threats identified. Data sources are standard and previously vetted by the NQF for the individual domains; missing data is exceedingly low per the report. 2b4. As above; the content validity assessment demonstrated that there were significant (and clinically meaningful differences) in all 4 domains between hospitals with 1 and 3 star ratings based on the composite measure. 2b5. I'm not entirely sure what this means 2b6. Missing data in the STS database appears to be quite rare across each of the four domains per the report.
- Yes; would appreciate examples from the database on policies to "maximize provider data completeness for elements that comprise the composite score"
- yes
- The conversion of score CIs into Star ratings is potentially problematic. Star ratings are assigned based on whether the score CI overlaps with the average performance. But two participants can have the same score and different star rankings based solely on the CI width (sample size in this case). It is entirely plausible that participants with high overall scores are assigned lower star ratings than participants who they do not statistically differ.
- No significant threats since using the STS database.
- No significant concerns regarding ability to identify meaningful differences with star rating identifying statistically significant better and worse performance compared to overall average. Comparability N/A. Greater than 5% missing data excluded therefore no concerns about threat to validity.
- not clear but rare and don't change results per developer
- Only concern is disparity in population represented

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?

- 2b2-3. Exclusions appear to be reasonable and justified for the IMA domain (anatomic considerations), medication on discharge process measures (in-house mortality)—and were previously endorsed by the NQF. All M&M cases are included without exclusions. Is this OK for the morbidity measure if a patient dies during the 30-day postoperative period? A hospital that has a high rate on in-house mortality could have a lower rate of 30-day morbidity simply because patients are not being exposed to the 30-day measurement period? The risk-adjustment approach appears to be comprehensive; mortality risk-adjustment has been previously vetted by the NQF and morbidity risk-adjustment was based on surgeon expert input and literature review – followed by a published analysis in 2009. 2b2. As above; Exclusions appear to be reasonable and justified for the IMA domain (anatomic considerations), medication on discharge process measures (in-house mortality)—and were previously endorsed by the NQF. All M&M cases are included without exclusions. Is this OK for the morbidity measure if a patient dies during the 30-day postoperative period? A hospital that has a high rate on in-house mortality could have a lower rate of 30-day morbidity simply because patients are not being exposed to the 30-day measurement period? 2b3. As above-- the risk-adjustment approach appears to be comprehensive; mortality risk-adjustment has been previously vetted by the NQF and morbidity risk-adjustment was based on surgeon expert input and literature review – followed by a published analysis in 2009.
- n/a
- Disagree
- No Issues with the exclusions or risk adjustment
- No other concerns/threats noted.
- Exclusions seem consistent with evidence and appropriate. Individual components are risk adjusted and align with conceptual description. Variables present at the start of care. Overall appropriate risk adjustment.
- acceptable
- No concerns

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?

- As above-- The weighting scheme seems to be very well documented and supported by published consensus statements. Furthermore, the content validity assessment demonstrated that there were significant (and clinically meaningful differences) in all 4 domains (including IMA %'s surprisingly) between hospitals with 1 and 3 stars. However, despite the content validity, the rules were developed from an "expert panel". I would be curious how this panel came to this relative weighting scheme to assign the final composite score (and star ratings). How were these weights developed? Cost associated with different events? Excess hospital days?
- n/a
- Agree
- Again, the weighting rules are such that the domains with the most variability have the lowest weight, making the overall performance distribution constrained
- See above, no significant concerns.
- Component measures fit the quality construct. The weighting rules are in alignment with expert assessment and empirical testing.

- moderate
- Yes

Criterion 3. Feasibility

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

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

- The data source for this measure is the STS registry. The STS database has more than 1,000 participants. Data are collected or generated and used by healthcare personnel during provision of care. Some institutions have full EHR capability; some may have partial or no availability. Some data elements are in defined fields in electronic sources and some must be abstracted. However, all data from participating institutions are submitted in electronic format following standard data specifications.
- STS registry participants pay annual fees of $3,500 to $4,750 depending on whether the majority of surgeons in the group are STS members. There is an additional fee of $150 per member and $350 per nonmember for surgeons listed on the database Participation Agreement. Most participants also purchase data entry software to submit data elements.

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?
- What effect is STS non-membership likely to have on feasibility; how significant would the effect be on measure use outcomes?

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—all data is collected as part of the STS database participation and the data collection has been previously vetted by the NQF for the individual domains.
- n/a
- Agree
- Feasibility has been established.
- This is already in practice/being used.
- All data elements are readily available and collected during care delivery. Some elements are not in electronic sources depending on the institution capability. This need for abstraction to convert to electronic submission to the STS database might cause increased burden on institutions without EHR capability, however not a significant impact to feasibility. Seems that STS membership is required for all registry participants, therefore non-membership would be a barrier.
- expense of belonging to STS; acceptable
- No concerns at this time

Criterion 4: Usability and Use

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

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

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

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

Current uses of the measure

Publicly reported? ☒ **Yes** ☐ **No**

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

OR

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

Accountability program details

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

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

- The developer states that each participant receives quarterly feedback reports providing a detailed analysis of the participant's performance including benchmarking. Participants also have access to a guide to help interpret performance results.

Feedback on the measure by those being measured or others

- The developer states that the STS Adult Cardiac Surgery Task Force meets periodically and provides input. The developer states they are working on developing real-time online dashboard-type reporting in response to requests from members.

Additional Feedback:

Questions for the Committee:

- How have the performance results been 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**

RATIONALE:

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

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

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

Improvement results

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

Star ratings in percentages, 2010-2018

Stars	2018	2017	2016	2015	2014	2013	2012	2011	2010
*	4.37	4.55	5.29	5.82	4.59	9.19	9.0	9.6	11.0
**	88.27	89.21	84.65	84.4	86.64	75.86	76.0	76.5	75.5
***	7.36	6.24	10.00	9.74	8.77	14.95	15.0	14.0	13.5

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

Unexpected findings (positive or negative) during implementation

None reported.

Potential harms

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

Additional Feedback:

Questions for the Committee:

- Are you aware of any unintended consequences related to this measure?
- Do the benefits of the measure outweigh any potential unintended consequences?

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

RATIONALE:

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?

- 4a1. Yes, through the STS public reporting website 4a2. Yes, all participants receive quarterly data reports with statistics to indicate performance relative to peers. An STS task force is charged with periodically reviewing the STS data product and providing feedback for the database and associated reports.
- n/a
- Agree
- The measure has been in use and publicly reported for years
- STS website/public reporting available.
- Measure is publicly reported and quarterly feedback to participants help to improve quality of care. Not in current or planned use for accountability programs. Input from the STS Adult Cardiac Surgery Taskforce provides real-world feedback from others.
- pass
- Already assessible, opportunity for patients & MD to obtain reports

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.

- 4b1. The usability of the individual domain performance scores is intuitive; the usability of the composite score is less obvious to me independent of its components 4b2. Unintended consequence: A hospital may be less motivated to improve its performance across a specific domain if their overall composite performance is "good enough". This goes back to my comments in 4b1 in terms of how the composite measure would actually motivate and/or drive QI/PI.
- n/a
- Agree
- Because the range of performance is very narrow, it is unclear how the results will motivate improvements. The Star rating system is also problematic because a site with higher performance might have fewer stars than a lower performing site (from which is isn't statistically different) depending on the width of their CIs
- Important measure of the overall quality of CABG.
- Improvements are shown with a reduction in variation. Not aware of any unintended consequences or additional potential harms other than noted risk aversion. Agree that benefits outweigh potential harm.
- no harms seen
- Drives medical team to be more intentional as there will be more transparency

Criterion 5: Related and Competing Measures

Related or competing measures

0114: Risk-Adjusted Postoperative Renal Failure
0115: Risk-Adjusted Surgical Re-exploration
0116: Anti-Platelet Medication at Discharge
0117: Beta Blockade at Discharge

0118: Anti-Lipid Treatment Discharge
0119: Risk-Adjusted Operative Mortality for CABG
0120: Risk-Adjusted Operative Mortality for Aortic Valve Replacement (AVR)
0121: Risk-Adjusted Operative Mortality for Mitral Valve (MV) Replacement
0122: Risk-Adjusted Operative Mortality for Mitral Valve (MV) Replacement + CABG Surgery
0123: Risk-Adjusted Operative Mortality for Aortic Valve Replacement (AVR) + CABG Surgery
0127: Preoperative Beta Blockade
0129: Risk-Adjusted Postoperative Prolonged Intubation (Ventilation)
0130: Risk-Adjusted Deep Sternal Wound Infection
0131: Risk-Adjusted Stroke/Cerebrovascular Accident
0134: Use of Internal Mammary Artery (IMA) in Coronary Artery Bypass Graft (CABG)
1501: Risk-Adjusted Operative Mortality for Mitral Valve (MV) Repair
1502: Risk-Adjusted Operative Mortality for Mitral Valve (MV) Repair + CABG Surgery
2514: Risk-Adjusted Coronary Artery Bypass Graft (CABG) Readmission Rate
2561 STS Aortic Valve Replacement (AVR) Composite Score
2563 STS Aortic Valve Replacement (AVR) + Coronary Artery Bypass Graft (CABG) Composite Score
2683: Risk-Adjusted Operative Mortality for Pediatric and Congenital Heart Surgery
3030 STS Individual Surgeon Composite Measure for Adult Cardiac Surgery
3031 STS Mitral Valve Repair/Replacement (MVRR) Composite Score
3032 STS MVRR + CABG Composite Score
3294 STS Lobectomy for Lung Cancer Composite Score
3534 30 Day All-cause Risk Standardized Mortality Odds Ratio following Transcatheter Aortic Valve Replacement (TAVR) [Currently in new measure endorsement process through Cardiovascular Standing Committee]

Harmonization

The 26 related measures identified are NQF-endorsed measures developed by or with STS, 11 of which are component measures of the CABG composite. The developer indicates they are harmonized.

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

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

- I'm unsure.
- n/a
- abstain
- NA
- Components of the composite measure are parts of other individual measures; there is some overlap with other STS measures.
- All related measures including component measures appear to be harmonized.
- no
- Quite a list of related/competing to review

Public and Member Comments

No comments have been submitted as of 01/07/2020.

Combined Methods Panel Scientific Acceptability Evaluation

Measure Number: 0696

Measure Title: STS CABG Composite Score

Type of measure:

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

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

Data Source:

☐ **Claims** ☐ **Electronic Health Data** ☐ **Electronic Health Records** ☐ **Management Data**
☐ **Assessment Data** ☐ **Paper Medical Records** ☐ **Instrument-Based Data** ☒ **Registry Data**
☐ **Enrollment Data** ☐ **Other**

Level of Analysis:

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

Measure is:

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

- **RELIABILITY: SPECIFICATIONS**

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

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

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

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

   Panel Member #1: Specifications were difficult to test using the information provided within the measure specification form. The full information exists within the included manuscripts and additional files. As specifications have not changed since last endorsement, I am assuming the materials submitted are satisfactory.

   Panel Member #2: No concerns. The composite specifications are well described.

   Panel Member #4: *None.*

   Panel Member #7: I have no concerns.

- **RELIABILITY: TESTING**

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

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

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

   Submission document: Testing attachment, section 2a2.2

   Panel Member #1: Methods seem adequate.

   Panel Member #2: The developer description of the reliability metric calculation is exemplary and a model for other developers. I would like to understand better whether low volume measured entities contribute equally to the reliability metric as high volume measured entities, or whether low volume measured entities contribute less than high volume measured entities.

   Panel Member #3: Reliability was assessed using a Bayesian implementation of the SNR. Overall reliability was 0.68, with a 95%. Credible interval of 0.63-0.73. Reliability for groups with N > 50 is 0.71. These values are acceptable.

   Panel Member #4: *Reliability for this composite measure was estimated using a method requiring MCMC simulation to estimate hospital-specific performance on the composite. The reliability measure was "defined as the estimated squared correlation between the set of hospital-specific estimates … and the corresponding unknown true values." I am not certain I fully understand this approach and how it could compare to potential alternative approaches.*

   Panel Member #5: A standard signal to noise analysis was conducted using 7-8 year old data. From the descriptive data given in the testing document for the older data vs the updated data presented in the MIF, it looks like overall performance on this composite has improved and the range of scores has narrowed. This suggests that the reliability analysis should be updated.

   Panel Member #6: *Bayesian MCMC signal-to-noise ratio calculation is appropriate.*

   Panel Member #7: The steward created a Markov Chain Monte Carlo simulation to estimate reliability. The simulation drew random samples from the Bayesian posterior probability distribution of each composite score estimate.

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

   Submission document: Testing attachment, section 2a2.3

   Panel Member #1: Results suggest the minimum threshold of number of operations per participant used as an inclusion criterion (≥10) is reasonable (r=0.68), with 50 or more cases per participant and measured year yielding above accepted reliability estimates (≥.71).

   Panel Member #2: The graph of "true score" vs. "measured score" is another exemplary method of interpreting the reliability metric results. The reporting of the reliability metric with alternative volume thresholds is useful, but a table that reports the reliability metric stratified at various volume categories would be preferred (and related to my question above about the contribution of low volume measured entities. My only concern is that the data used for testing (2013-2014) is not the same as the data used for reporting (2016-2017) and there is reason to believe that the reliability metric might be less in more recent data given the narrow range of scores (0.91-98).

The reliability metric value of 0.68 demonstrates "substantial" relialbity on the Landis scale and is a more credible result than the Beta-Binomial calculations that I am convinced are too high.

Panel Member #3: Reliability was assessed using a Bayesian implementation of the SNR. Overall reliability was 0.68, with a 95%. Credible interval of 0.63-0.73. Reliability for groups with N > 50 is 0.71. These values are acceptable.

Panel Member #4: *Reliability estimates were mean of 0.68, ranging from 0.71 in participants with 50 or more operations (863 of 1024 participants) and 0.72 for those with 100 or more operations (582 or 1024 participants). I am curious how the reliability of the composite compares to the reliability of its components.*

Panel Member #5: Overall reliability is adequate and, as expected, better for units with higher sample sizes. As stated above, the data used for these analyses may not accurately reflect current performance.

Panel Member #6: *Posterior mean of reliability is 0.68, which is ok, but not high.*

Panel Member #7: Mean reliability was 0.68, with a 95% posterior credible interval ranging from 0.63 to 0.73. Higher surgery volume was associated with modestly higher reliability.

8. Was the method described and appropriate for assessing the proportion of variability due to real differences among measured entities? NOTE: If multiple methods used, at least one must be appropriate.

   Submission document: Testing attachment, section 2a2.2

   ☒ **Yes**

   ☐ **No**

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

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

   Submission document: Testing attachment, section 2a2.2

   ☐ **Yes**

   ☐ **No**

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

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

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

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

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

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

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

    Panel Member #1: I have no concerns with the reliability testing and results, other than the score level reliability being moderate (0.68), thus the moderate rating.

    Panel Member #2: The description, methods, interpretation and results are exemplary but the data are becoming dated and may not have been updated since the most recent endorsement maintenance. Reliability is not a static characteristic of the measure but a context specific characteristic, and part of that context is trends in performance from 2014 to 2019.

Panel Member #3: Reliability was assessed using a Bayesian implementation of the SNR. Overall reliability was 0.68, with a 95%. Credible interval of 0.63-0.73. Reliability for groups with N > 50 is 0.71. These values are acceptable.

Panel Member #4: *To restate: I am not certain I fully understand this approach and how it could compare to potential alternative approaches. The rating provided may be optimistic. I will appreciate learning others' perspectives.*

Panel Member #5: See comments above. If the overall performance has increased and the range has narrowed, the reliability analysis based on older data may not reflect current reality.

Panel Member #6: *Posterior mean of reliability is 0.68, which is ok, but not high.*

Panel Member #7: Reliability is modest. Clearly, as with any composite metric, the reliability is driven by the reliability of individual components (and their respective weights).

- **VALIDITY: ASSESSMENT OF THREATS TO VALIDITY**

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

Submission document: Testing attachment, section 2b2.

Panel Member #1: No concerns

Panel Member #2: No concerns. The exclusions are well justified and empirically material.

Panel Member #3: none

Panel Member #4: *None.*

Panel Member #5: No Concerns

Panel Member #6: *The detailed analysis seems to show that the exclusions matter a lot. The developers suggest that this is ok because the exclusions are appropriate and necessary (which is a matter of face validity), but this has not been formally assessed.*

Panel Member #7: I have no concerns.

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

Submission document: Testing attachment, section 2b4.

Panel Member #1: I have no concerns. To enable an assessment of improvement over time, it would be interesting to see how scores from the same participants change over time. I assume that in the data provided participants may change from period to period.

Panel Member #2: The only concern is the use of data from 2013-2014. If the trends demonstrated in 2b4.2 have continued to 2019 then all measured entities would be two stars.

Panel Member #3: none

Panel Member #4: *No significant concerns. Past measures have moderate ability to predict future performance. Performance differences may be considered meaningful.*

Panel Member #5: The range of performance is between ~92% and 98%. The reliability of the measure has not been updated. It therefore hard to know if meaningful differences can be detected.

Panel Member #6: *The developers use a star system based on a Bayesian model to identify approximately 20% of the participants that are significantly above or below average. While this is statistically appropriate, it says nothing about meaningful differences. Given that the overall scores are so close to 100%, whether these differences are clinically meaningful is questionable.*

Panel Member #7: The methodology of the measure facilitates not only a composite measure value, but also an associated 95% credible interval, thus permitting inference about differences between any two participants.

14. **Please describe any concerns you have regarding comparability of results if multiple data sources or methods are specified.**
Submission document: Testing attachment, section 2b5.
Panel Member #2: No concerns. Not applicable.
Panel Member #3: none
Panel Member #4: *None.*
Panel Member #5: NA
Panel Member #6: *No concerns.*
Panel Member #7: This item is not applicable.

15. **Please describe any concerns you have regarding missing data.**
Submission document: Testing attachment, section 2b6.
Panel Member #1: No concerns. Overall missing data were very rare.
Panel Member #2: No concerns. Measured entities with more than 5% missing data were excluded; otherwise measures were imputed with the negative value (incentivizing participants to report). Empirically missing data is immaterial.
Panel Member #3: none
Panel Member #4: *None.*
Panel Member #5: None
Panel Member #6: *No concerns.*
Panel Member #7: I have 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**

Panel Member #1: No concerns. Information on the risk-adjustment model for the Absence of Operative Mortality domain was not provided as part of this submission but included in a separate endorsed NQF measure (0119).

Information provided for the morbidity domain risk-adjustment model was satisfactory.

Panel Member #2: The two outcome domains (mortality and morbidity) are risk-adjusted and individually endorsed.

Panel Member #3: Used backward stepwise selection for risk factor selection. This is not ideal and can result in the omission of clinically important risk factors. Model was cross-validated after splitting data into development and validation data set. C statistic for composite mortality/morbidity model is 0.73, which is acceptable. Calibration plot indicated excellent calibration.

This measure is based on data that is now at least 5 years old. The risk adjustment model needs to be updated to reflect more contemporary data. For the individual risk prediction model validation section, this is based on data that is > 10 years old published in 2009.

Panel Member #4: *Robust.*

Panel Member #5: Individual component measures are NQF endorsed. Calibration and discrimination are adequate.

Panel Member #6: *Well developed and tested.*

Panel Member #7: Individual components of the composite are risk-adjusted for an array of important clinical factors.

<u>For cost/resource use measures ONLY</u>:

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

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

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

- **VALIDITY: TESTING**

19. **Validity testing level:** ☒ **Measure score** ☐ **Data element** ☐ **Both**

20. **Method of establishing validity of the measure score:**

    ☒ **Face validity**

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

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

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

    Submission document: Testing attachment, section 2b2.2

    Panel Member #1: Face validity was supported.

    I found the other validity testing challenging.

    Empirical validity was assessed using predictive validity/longitudinal stability, i.e., associations of scores over different time periods. An association with a different construct that is expected to be correlated with measure 0696 was not assessed. A longitudinal prediction may support the validity of the scoring system, or validity/stability of a risk-adjustment model. Otherwise, stability of score over time is expected as long as performance does not change over time. Hopefully, performance can improve over time, decreasing this type of longitudinal predictive validity, without decreasing its empirical validity. Could the developers demonstrate that the scores are associated with another related measure? If for some reason this is not possible, it would be helpful to understand why.

A demonstration that the measure scores differ between patient groups in clinical logical and expected ways would help support known-groups validity.

Content validity was tested using associations between score components (the 4 domain scores) and the composite performance score (3 star scoring) during the same measurement period. A positive association is expected to exist as the components create the composite score. I found this test to be somewhat circular.

Panel Member #2: The developer uses two approaches to assess validity: 1) construct validity by demonstrating an implicit quality construct and 2) content validity by demonstrating the component measures move in the same direction as the composite. The first approach merely demonstrates that the implicit QC has a persistent component which is supportive regarding use of the composite in comparative reporting (since current decisions are based on historical data).

Panel Member #3: Assessed the predictive validity by examining the correlation between performance on the Composite measure between 2 different time periods. This comparison, again, is based on data that is 6 years old and needs to be updated. The Pearson correlation coefficient is 0.74, which is acceptable.

Panel Member #4: *Methods to assess validity are reasonable.*

Panel Member #5: "We tested the predictive validity of the composite measure. Predictive validity means that the results of this measure are predictive of future performance. We assessed the extent to which performance on the STS composite measure remains stable over time. In other words, does the composite score performed at one point in time accurately predict performance at some later time?" This is an analysis of stability not validity. Predictive validity usually assesses the link between process measure performance and subsequent outcomes.

Panel Member #6: *The developers assert that face validity implies that the measure is regarded as useful and valid by its intended users, including providers, consumers, payers, and regulators, and report "near-universal acceptance of this measure by all stakeholders", and its use by Consumer Reports, for nearly 5 years. This is not the systematic and transparent process expected.*

What the developers describe as content validity ("that the four domains in the composite are broadly representative of the latent construct 'isolated CABG quality'") is a form of face validity. However, the developers do not describe any formal process for assessing it (as face validity).

Empirical testing consists of assessing whether the composite score calculated at one point in time accurately predict performance at some later time, or in other words, whether the STS composite measure remains stable over time. Since it is the same measure, this is reliability rather than validity.

The developers also examine the 4 components of the composite vary according to the star rating (see Q #12) in both the previous and current measurement year. This essentially demonstrates that the 4 components are each correlated with the composite (and indirectly with each other, but it does not really establish validity.

Panel Member #7: The steward appealed to empirical validity testing, face validity, and content validity. Regarding empirical validity, the steward assessed year-over-year changes in the composite measure, as well as the relationship between the composite measure in one year and values of its components in a subsequent year. Regarding content validity, the steward assessed whether star ratings (derived from the composite measure) were correlated with individual domains of the composite.

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

Submission document: Testing attachment, section 2b2.3

The concerns described above question the ability of the validity testing results to support score level known groups or empirical validity.

Panel Member #2: The results demonstrate moderate validity or validity maturity level I (an empirical association between the implicit QC and the material outcome).

Panel Member #3: Assessed the predictive validity by examining the correlation between performance on the Composite measure between 2 different time periods. This comparison, again, is based on data that is 6 years old and needs to be updated. The Pearson correlation coefficient is 0.74, which is acceptable.

Panel Member #4: *Face validity, content validity, and empirical validity.*

Panel Member #5: I don't think the empirical testing analyses (required for maintenance) pertains to validity, rather I think it is testing lagged test-retest reliability (stability).

Panel Member #6: *Since the definition of face validity is not what NQF expects (see Q. #20), the results are not convincing.*

The correlation between the score at one time and another is relatively low (0.63 or 0.64), which together with the fact that it is not an appropriate measure of validity (see Q #20), is not convincing.

As noted in Q #20, the comparison of the score components with the start rating does not establish the validity of the composite.

Panel Member #7: The analyses supported validity by demonstrating positive serial correlation in the composite measure and positive correlations between the composite measure and its domains.

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

    Submission document: Testing attachment, section 2b1.

    ☒ **Yes**

    ☒ **No**

    Panel Member #1: I believe that addressing the concerns above could solve this issue.

    ☐ **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 <u>are</u> threats to validity and/or relevant threats to validity were <u>not assessed OR</u> 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 <u>is required</u>; if not conducted, should rate as INSUFFICIENT.)

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

Panel Member #1: As noted above, additional information on empirical validity and/or know groups validity would enable a validity rating.

Panel Member #2: Although the results demonstrate moderate validity the data used to demonstrate validity are from 2013-2014. Validity is not a static characteristic of the measure but rather a context dependent characteristic, and part of that context is intended use. Therefore the threat to validity in 2019 is changes to the context since 2014 that may impact the validity assessment.

Panel Member #3: Assessed the predictive validity by examining the correlation between performance on the Composite measure between 2 different time periods. This comparison, again, is based on data that is 6 years old and needs to be updated. The Pearson correlation coefficient is 0.74, which is acceptable.

- **Panel Member #4:** *This is a well-established and thoughtfully-crafted measure.*

Panel Member #5: Reliability testing not updated to reflect the new distribution, therefore it is hard to tell if meaningful differences can be detected in the current compressed performance range. Empirical validity testing assessed year-over-year stability not validity.

Panel Member #6: *As discussed above, neither the methods to assess face validity nor the empirical testing were appropriate.*

Panel Member #7: The associations between the composite measure and its components are strong.

- **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**

    Panel Member #1: I have a couple of clarification questions:

    1. I am assuming the 4 domains are expected to be correlated. If so, inter-domain correlations or some type of factor analyses would be informative to better understand how domains are associated to one another.
    2. Were the correlations tested between each domain and the composite score and between each domain and its components, item-rest-correlations or simple item-total test score correlations? Item-rest correlations would be more informative on how each domain or component is associated with its composite.

    Weights selected for each domain were adequately supported.

    Since this measure has already been endorsed, I'm rating this as moderate. For a new composite I would have probably rated this as insufficient. Please advise if this is reasonable.

    Panel Member #2: The weight that is applied to a given domain is inversely proportional to the standard deviation of the domain-specific scores. There is no conceptual rational for this weighting scheme. There is no judgement about how the weight addresses the importance, reliability, validity, or usability of the component measures. A conceptual rational would be justified based on competing or uncertain importance. For example the components might be weighted to the degree that the component explains a target outcome.

    Panel Member #3: The weighting of the components was established in collaboration with a technical expert panel. By definition, these weights are expert-based.

Panel Member #4: *The components correlate with*

"The composite score is a weighted average across four domains. 81% of the total weight was assigned to mortality, 10% to morbidity, 7% to IMA and 3% to medications." This weighting reflects expert opinion and was evaluated empirically to demonstrate that a 1% change in mortality would impact composite score as would an 8% change in morbidity.

Panel Member #5: I did not see an analysis of the composite score construction.

Panel Member #6: *As explained in the testing report, the aggregation and weighting make sense both in terms of expert assessment and empirical testing.*

Panel Member #7: Each component is positively correlated with the composite score. However, the composite score takes 81% of its weight from mortality and 10% of its weight from morbidity.

- **ADDITIONAL RECOMMENDATIONS**

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

Panel Member #1: I believe the SMP should discuss these concerns before forwarding them to the standing committee.

Panel Member #2: The scientific acceptability of this composite measure in 2014 was high to moderate. The scientific acceptability in 2019 is not possible to determine based on the data provided except to assume no change in context.

Panel Member #3: The composite measure is based on data that is now at least 5 years old. The risk adjustment model needs to be updated to reflect more contemporary data. For the individual risk prediction model validation section, this is based on data that is > 10 years old published in 2009.

I do not see any information on the methodology used to combine the component scores into the composite scores. This information should be supplied by the measure developer. It would also be useful to see a data dictionary in order to assess data specification.

Panel Member #4: *No.*

Panel Member #5: The conversion of score CIs into Star ratings is potentially problematic. Star ratings are assigned based on whether the score CI overlaps with the average performance. But two participants can have the same score and different star rankings based solely on the CI width (sample size in this case). It is entirely plausible that participants with high overall scores are assigned lower star ratings than participants who they do not statistically differ.

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

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

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

0696_-_NQF_evidence_attachment_v7_1-112619update.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)

Measure Number (*if previously endorsed*)**:** 0696

Measure Title: STS CABG Composite Score

IF the measure is a component in a composite performance measure, provide the title of the Composite Measure here: Click here to enter composite measure #/ title

Date of Submission: 11/15/2019

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

Outcome

☐ Outcome:

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

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

☐ Intermediate clinical outcome (*e.g., lab value*): Click here to name the intermediate outcome

☐ Process:

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

☐ Structure: Click here to name the structure

☒ Composite: Four domains: 1. Operative Mortality (outcome); 2. Postoperative Major Morbidity (outcome); 3. IMA use (process); 4. Perioperative Medications (process). See details under 1a.2 below and in Appendix.

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 is the single most important negative outcome associated with a surgical procedure. The published literature on coronary artery bypass grafting, including the references provided below, is full of examples of services/care processes that impact operative mortality. Pre-operative 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.

Major Morbidity

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. Because it remains a major source of morbidity and mortality for the CABG population, deep sternal wound infection rate is a relevant quality parameter and an important performance metric for a cardiac surgery program. 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.

Modalities to decrease the rate of prolonged intubation include physician-supervised protocols for extubation implemented by nurses and respiratory therapists, improved pre-operative preparation of patients, reduction of postoperative bleeding, and intra-operative protocolized anesthesia care. Current implementation of these modalities 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, and lower rates of operative mortality.

Modalities to decrease the rate of postoperative stroke include measures to maintain blood pressure and perfusion, avoidance of atrial fibrillation, anticoagulation protocols, etc.

Identification of clinical precursors of postop 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), postop kidney injury can be significantly reduced.

Reoperation for bleeding can be reduced by careful intraoperative hemostasis, use of topical and systemic hemostatic agents, and rapid achievement of postoperative normothermia.

Internal Mammary Artery

The internal mammary artery has definitively and repeatedly been shown to be the best conduit for coronary bypass grafting. It has been shown to have the highest patency rates compared to other conduits and its use substantially prolongs patient survival in the long term over other conduit choices.

Perioperative Medications

Preoperative Beta Blockade

The most compelling justification for preoperative beta blockade use, and its inclusion as a performance measure for cardiac surgery, is its impact on the development of postoperative atrial fibrillation. This common complication occurs in about 23% of patients undergoing isolated CABG surgery by STS Database participants, and it results in increased resource utilization (LOS). The Virginia Cardiac Surgery Quality Initiative (VCSQI) found that atrial fibrillation added an average 10.3% ($2,744) and 2.2 days length of stay to a typical isolated CABG hospitalization. Postoperative atrial fibrillation increases the risk of stroke, an often devastating complication, as well as other thromboembolic complications. It may produce hemodynamic compromise in some patients and at the very least is symptomatically unpleasant. Multiple studies show that the development of postoperative atrial fibrillation is an independent predictor of worse long-term survival following CABG surgery.

Beta Blockade at Discharge

The use of postoperative beta blockers is known to protect patients both at one year and long term (greater than 5 years) from death following cardiac surgery. This effect is associated with a 46% risk reduction in death at one year and 35% risk reduction in mortality during long-term follow-up.

Anti-Lipid Treatment at Discharge

Recognizing the importance of statins, including their protean and lipid-lowering effects, patients with bypass conduits have the potential to benefit from anti-lipid therapy as noted in the references. In addition, a high level of evidence was found that statins reduce total mortality in individuals with a history of prior events related to atherosclerotic cardiovascular disease.

Anti-Platelet Medication at Discharge

The provision of anti-platelet therapy at discharge is currently accepted as standard of care for promotion of secondary prevention of coronary artery disease and improved long-term vein graft patency. It is expected that long-term mortality reduction following CABG surgery will occur with ASA therapy. Multiple peer review publications listed below provide evidence for this marker.

References:

- 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. Explaining Racial Disparities in Outcomes After Cardiac Surgery: The Role of Hospital Quality. JAMA Surg. 2014;149(3):223-7.
- 1999. ASHP Therapeutic Guidelines on Antimicrobial Prophylaxis in Surgery. American Society of Health-System Pharmacists. Am J Health Syst Pharm 56: 1839-88
- Edwards FH, Engelman RM, Houck P, Shahian DM, Bridges CR. 2006. The Society of Thoracic Surgeons Practice Guideline Series: Antibiotic Prophylaxis in Cardiac Surgery, Part I: Duration. Ann Thorac Surg 81: 397-404
- Tamayo E, Gualis J, Florez S, Castrodeza J, Bouza JM, Alvarez FJ. 2008. Comparative study of single-dose and 24-hour multiple-dose antibiotic prophylaxis for cardiac surgery. J Thorac Cardiovasc Surg 136: 1522-7
- Engelman R, Shahian D, Shemin R, Guy TS, Bratzler D, Edwards F, Jacobs M, Fernando H, Bridges C. 2007. The Society of Thoracic Surgeons practice guideline series: Antibiotic prophylaxis in cardiac surgery, part II: Antibiotic choice. Ann Thorac Surg 83: 1569-76
- Gupta A, Hote MP, Choudhury M, Kapil A, Bisoi AK. 2010. Comparison of 48 h and 72 h of prophylactic antibiotic therapy in adult cardiac surgery: a randomized double blind controlled trial. J Antimicrob Chemother 65: 1036-41
- Harbarth S, Samore MH, Lichtenberg D, Carmeli Y. 2000. Prolonged antibiotic prophylaxis after cardiovascular surgery and its effect on surgical site infections and antimicrobial resistance. Circulation 101: 2916-21
- Furnary AP, Zerr KJ, Grunkemeier GL, Starr A. Continuous intravenous insulin infusion reduces incidence of deep sternal wound infection in diabetic patients after cardiac surgical procedures. *Ann Thorac Surg* Feb 1999; 67:352-362.
- Lazar HL, McDonnell M, Chipkin SR, Furnary AP, Engelman RM, Sadhu AR, Bridges CR, Haan CK, Svedjeholm R, Taegtmeyer H, Shemin RJ; Society of Thoracic Surgeons Blood Glucose Guideline Task Force. The Society of Thoracic Surgeons practice guideline series: Blood glucose management during adult cardiac surgery. Ann Thorac Surg. 2009 Feb;87(2):663-9.
- 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.
- 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.
- 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.
- 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.
- Shroyer LA, 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.
- 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.
- 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.

- 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.
- Kramer RS, Quinn RD, Groom RC, Braxton JH, Malenka DJ, Kellett MA, Brown JR for the Northern New England Cardiovascular Disease Study Group. Same admission cardiac catheterization and cardiac surgery: is there an increased incidence of acute kidney injury? Ann Thorac Surg 2010;90:1418-1424.
- Mangano CM, Diamondstone LS, Ramsay JG, et al. Renal dysfunction after myocardial revascularization: risk factors, adverse outcomes, and hospital resource utilization: the Multicenter Study of Perioperative Ischemia Research Group. *Ann Intern Med.* 1998;128(3):194-203.
- Ranucci M, Ballotta A, Agnelli B, Frigiola A, Mencanti L, Castelvecchio S, for the Surgical and Clinical Outcome Research (SCORE) Group. Acute kidney injury in patients undergoing cardiac surgery and coronary angiography on the same day. Ann Thorac Surg 2103;95:513-519.
- Rosner MH, Okusa MD. Acute kidney injury associated with cardiac surgery. Clin J Am Soc Nephrol 2006;1:19-32.
- Shahian DM, Edwards FH, Ferraris VA, Haan CK, Rich JB, Normand SLT, DeLong ER, O'Brien SM, Shewan CM, Dokholyan RS, Peterson ED. Quality Measurement in adult cardiac surgery: Part 1-conceptual framework and measure selection. Ann Thorac Surg 2007;83:S3-S12
- 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.
- 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.
- 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.
- 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.
- Abramov D, Tamariz MG, Sever JY, Christakis GT, Bhatnagar G, Heenan AL, Goldman BS, Fremes SE. The influence of gender on the outcome of coronary artery bypass surgery. *Ann Thorac Surg.* 2000;70:800-806.
- Arkansas Foundation for Medical Care. *Coronary Artery Bypass Graft Surgery: Performance Measures and Risk Adjustment Methodology.* Final Report to the Centers for Medicare and Medicaid Services; September 2002.
- Ferguson TB Jr, Coombs LP, Peterson ED. Internal thoracic artery grafting in the elderly patient undergoing coronary artery bypass grafting: room for process improvement? *J Thorac Cardiovasc Surg.* 2002;123(5):869-880.
- Leavitt B, O'Connor GT, et al. Use of the internal mammary artery graft and in-hospital mortality and other adverse outcomes associated with coronary artery bypass surgery. *Circulation.* 2001;103(4):507-512.
- Morris RJ, Strong MD, et al. Internal thoracic artery for coronary artery grafting in octogenarians. *Ann Thorac Surg.* 1996;62:16-22.
- Loop FD, Lytle BW, Cosgrove DM, et al. Influence of the internal-mammary-artery graft on 10-year survival and other cardiac events. *N Engl J Med*. 1986 Jan 2;314(1):1-6.
- Lytle BW, Blackstone EH, Loop FD, et a. Two internal thoracic artery grafts are better than one. *J Thorac Cardiovasc Surg*. 1999 May;117(5):855-72.
- Speir AM, Kasirajan V, Barnett SD, Fonner E Jr. Additive costs of postoperative complications for isolated coronary artery bypass grafting patients in Virginia. Ann Thorac Surg 2009 Jul;88(1):40-5.

- D'Agostino RS, Svensson LG, Neumann DJ, Balkhy HH, Williamson WA, Shahian DM. Screening carotid ultrasonography and risk factors for stroke in coronary artery surgery patients. Ann Thorac Surg 1996 Dec;62(6):1714-23.
- Likosky DS, Leavitt BJ, Marrin CA, Malenka DJ, Reeves AG, Weintraub RM, et al. Intra- and postoperative predictors of stroke after coronary artery bypass grafting. Ann Thorac Surg 2003 Aug;76(2):428-34.
- Stamou SC, Hill PC, Dangas G, Pfister AJ, Boyce SW, Dullum MK, et al. Stroke after coronary artery bypass: incidence, predictors, and clinical outcome. Stroke 2001 Jul;32(7):1508-13.
- Villareal RP, Hariharan R, Liu BC, Kar B, Lee VV, Elayda M, et al. Postoperative atrial fibrillation and mortality after coronary artery bypass surgery. J Am Coll Cardiol 2004 Mar 3;43(5):742-8.
- Mariscalco G, Klersy C, Zanobini M, Banach M, Ferrarese S, Borsani P, et al. Atrial fibrillation after isolated coronary surgery affects late survival. Circulation 2008 Oct 14;118(16):1612-8.
- El-Chami MF, Kilgo P, Thourani V, Lattouf OM, Delurgio DB, Guyton RA, et al. New-onset atrial fibrillation predicts long-term mortality after coronary artery bypass graft. J Am Coll Cardiol 2010 Mar 30;55(13):1370-6.
- Filardo G, Hamilton C, Hebeler RF, Jr., Hamman B, Grayburn P. New-onset postoperative atrial fibrillation after isolated coronary artery bypass graft surgery and long-term survival. Circ Cardiovasc Qual Outcomes 2009 May;2(3):164-9.
- Bramer S, van Straten AH, Soliman Hamad MA, Berreklouw E, Martens EJ, Maessen JG. The impact of new-onset postoperative atrial fibrillation on mortality after coronary artery bypass grafting. Ann Thorac Surg 2010 Aug;90(2):443-9.
- Crystal E, Connolly SJ, Sleik K, et al. Interventions on prevention of postoperative atrial fibrillation in patients undergoing heart surgery: a meta-analysis. *Circulation.* 2002;106(1):75-80.
- Kim MH, Deeb GM, Morady F, et al. Effect of postoperative atrial fibrillation on length of stay after cardiac surgery (The Postoperative Atrial Fibrillation in Cardiac Surgery study [PACS(2)]). *Am J Cardiol.* 2001;87(7):881-885.
- Maisel WH, Rawn JD, Stevenson WG. Atrial fibrillation after cardiac surgery. *Ann Intern Med.* 2001;135(12):1061-1073.
- Villareal RP, Hariharan R, Liu BC, et al. Postoperative atrial fibrillation and mortality after coronary artery bypass surgery. *J Am Coll Cardiol.* 2004;43(5):742-748.
- 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.
- Charlson ME, Isom OW. Care after coronary-artery bypass surgey. *N Engl J Med*. 2003;348:1456-63.
- Chen J, Radford MJ, Wang Y, Marciniak TA, Krumholz HM. Are beta-blockers effective in elderly patients who undergo coronary revascularization after acute myocardial infarction? *Arch Intern Med*. 2000;160:947-52.
- Chan AYM, McAlister FA, Norris, CM, et al. Effect of B-Blocker use on outcomes after discharge in patients who underwent cardiac surgery. *J Thorac Cardiovasc Surg*. 2010;140:182-7.
- Stone NJ, Robinson JG, Lichtenstein AH, et al. 2013 ACC/AHA guideline on the treatment of blood cholesterol to reduce atherosclerotic cardiovascular risk in adults: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. J Am Coll Cardiol. 2014;63:2889-934.
- Campeau L. Lipid Lowering and coronary bypass graft surgery. *Curr Opin Cardiol. 2*000;15(6):395-399.
- Denton TA, Fonarow GC, LaBresh KA, et al. Secondary prevention after coronary bypass: the American Heart Association "Get with the Guidelines" program. *Ann Thorac Surg.* 2003;75(3):758-760.
- Faulkner MA, Wadibia EC, Lucas BD, et al. Impact of pharmacy counseling on compliance and effectiveness of combination lipid-lowering therapy in patients undergoing coronary artery revascularization: a randomized, controlled trial. *Pharmacotherapy.* 2003;20(4):410-416.
- Hunninghake DB. Is aggressive cholesterol control justified? review of the post-coronary artery bypass graft trial. *Am J Cardiol.* 1998;82(10B):45T-48T.
- 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.

- Shah SJ, Waters DD, Barter P, Kastelein JJP, Shepherd J, Wenger NK, DeMicco DA, Breazna A, LaRosa JC. Intensive Lipid-lowering with Atorvastatin for secondary prevention in patients after coronary artery bypass surgery. J Am Coll Cardiol.2008; 51(20):1938-1943.
- Vaduganathan M, Stone NJ, Lee R, McGee EC, Malaisrie SC, Silverberg RA, McCarthy,PM. Perioperative statin therapy reduces mortality in normolipidemic patients undergoing cardiac surgery. J Thorac Cardiovasc Surg 2010;140(5):1018-1027
- Foody JM, Ferdinand FD, Galusha D, et al. Patterns of secondary prevention in older patients undergoing coronary artery bypass grafting during hospitalization for acute myocardial infarction. *Circulation.* 2003;108(Suppl-1):II24-II28.
- Mangano DT. Aspirin and mortality from coronary bypass surgery. *N Engl J Med.* 2002;347(17):1309-1317.
- Topol EJ. Aspirin with bypass surgery. *N Engl J Med.* 2002;347(17):1359-1360.
- 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.
- Hillis DL, Smith PK, Anderson JL, et al. 2011 ACCF/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. J Am Coll Cardiol 2011;58:2584-614.
- Ferraris VA, Saha SP, Oestreich JH, et al. 2012 update to the Society of Thoracic Surgeons guideline on use of antiplatelet drugs in patients having cardiac and noncardiac operations. Ann Thorac Surg 2012;94:1761-81.
- Sousa-Uva M, Storey R, Huber K, et al. Expert position paper on the management of antiplatelet therapy in patients undergoing coronary artery bypass graft surgery. Eur Heart J 2014;35:1510-14.
- Ferraris VA, Bolanos MD. Use of antiplatelet drugs after cardiac operations. Semin Thoracic Surg 2014; 26:223-230.
- Deo SV, Dunlay SM, Shah IK, et al. Dual anti-platelet therapy after coronary artery bypass grafting; is there any benefit? A systematic review and meta-analysis. J Card Surg 2013;28:109-16.
- Ebrahimi R, Bakaeen FG, Uberoi A, et al. Effect of clopidogrel use post coronary artery bypass surgery on graft patency. Ann Thorac Surg 2014;97:15-21.
- Kim DH, Daskalakis C, Silvestry SC, et al. Aspirin and clopidogrel use in the early postoperative period following on-pump and off-pump coronary artery bypass grafting. J Thorac Cardiovasc Surg 2009;138:1377-84.

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

Source of Systematic Review: • **Title** • **Author** • **Date** • **Citation, including page number** • **URL**	
Quote the guideline or recommendation verbatim about the process, structure or intermediate outcome being measured. If not a guideline, summarize the conclusions from the SR.	
Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	
Provide all other grades and definitions from the evidence grading system	
Grade assigned to the **recommendation** with definition of the grade	
Provide all other grades and definitions from the recommendation grading system	
Body of evidence: • Quantity – how many studies?	

• Quality – what type of studies?	
Estimates of benefit and consistency across studies	
What harms were identified?	
Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	

1a.4 OTHER SOURCE OF EVIDENCE

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

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 isolated CABG in two consecutive time periods, July 2015 – June 2016 and July 2016 – June 2017.

The table below summarizes the distributions of the STS CABG composite score in the last four quarterly harvests for which the composite scores were calculated. The fall harvests cover data from July of the previous year until June of the current year. The spring harvests cover data in the previous calendar year.

Distribution of STS isolated CABG composite measure in the latest four STS harvests for which the measure was reported

Stat	STS Harvests*			
	Latest	Spring 2017	Fall 2016	Spring 2016
# Participant	945	1006	882	1026
# Operations	145815	150882	129972	149917
Mean	0.967	0.967	0.967	0.966
STD	0.00972		0.0109	0.0102	0.0104
IQR	0.0123	0.0142	0.0131	0.0134
Percentiles				
0%	0.919	0.923	0.917	0.912
10%	0.954	0.952	0.954	0.953
20%	0.959	0.958	0.960	0.958
30%	0.962	0.962	0.964	0.962
40%	0.965	0.965	0.966	0.965
50%	0.968	0.968	0.969	0.968
60%	0.970	0.971	0.971	0.970
70%	0.972	0.973	0.974	0.973
80%	0.975	0.976	0.976	0.975
90%	0.978	0.980	0.978	0.978
100%	0.985	0.989	0.986	0.986
US Geographic Region				
Midwest		267	283	261	290
Northeast		126	129	112	134
South	359	381	327	386
West	185	207	178	216
Other	8	6	4	0

* Composite measure analysis of each harvest uses the most recent one year of data until the end of last quarter. For example spring 2017 harvest uses data until December 2016.

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

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 the STS participant (typically a hospital, a hospital group, or a surgeon group). It is not a patient or operation level measure. Therefore at the composite score level, we

do not provide data stratified by patient characteristics. Instead, we provide results stratified by participant characteristics.

Distribution of isolated CABG composite measures by regions, fall 2017 harvest, July 2016 - June 2017

Stat	Midwest		Northeast		South	West	Other*
# Participant	267	126	359	185	8		
# Operations	33448	24800	63107	22601	1859		
Mean	0.967	0.970	0.966	0.966	0.957		
STD	0.00932		0.00848		0.0101	0.00958	0.0073
IQR	0.0109	0.0114	0.0119	0.0128	0.012		
Percentiles							
0%	0.919	0.945	0.931	0.937	0.945		
10%	0.957	0.958	0.951	0.953	0.948		
20%	0.961	0.963	0.958	0.958	0.950		
30%	0.964	0.967	0.962	0.961	0.952		
40%	0.966	0.970	0.964	0.964	0.955		
50%	0.969	0.972	0.966	0.967	0.958		
60%	0.971	0.974	0.969	0.969	0.960		
70%	0.973	0.976	0.971	0.972	0.962		
80%	0.975	0.977	0.974	0.974	0.964		
90%	0.978	0.979	0.978	0.977	0.964		
100%	0.983	0.984	0.984	0.985	0.965		

*Non-North American/Canadian

Distribution of isolated CABG composite measures by regions, fall 2016 harvest, July 2015- June 2016

Stat	Midwest		Northeast		South	West	Other*
# Participant	261	112	327	178	4		
# Operations	32530	20875	55513	20427	627		
Mean	0.968	0.971	0.966	0.965	0.960		
STD	0.00911		0.00734		0.0109	0.0114	0.00938
IQR	0.0128	0.00887		0.0141	0.0144	0.00922	
Percentiles							
0%	0.928	0.943	0.927	0.917	0.946		
10%	0.956	0.963	0.952	0.952	0.950		
20%	0.961	0.965	0.959	0.957	0.954		
30%	0.964	0.968	0.962	0.961	0.958		
40%	0.967	0.969	0.965	0.965	0.961		
50%	0.969	0.972	0.968	0.968	0.962		
60%	0.971	0.974	0.971	0.970	0.964		
70%	0.974	0.975	0.973	0.972	0.965		
80%	0.975	0.978	0.976	0.974	0.966		

90%	0.978	0.980	0.979	0.977	0.966
100%	0.984	0.986	0.985	0.985	0.967

*Non-North American/Canadian

(If data in above tables does not display clearly, please see versions in Appendix.)

At the individual domain level, the risk-adjusted odds ratio associated with sex and race were:

Risk-adjusted odds ratio of mortality:

- Female (at BSA=1.8) v male (at BSA=2.0): 1.59 (95% confidence interval: 1.45-1.74)
- Black v white (including patients with race other than white, black, Asian): 1.17 (1.03 – 1.32)
- Asian v white (including patients with race other than white, black, Asian): 0.97 (0.80 – 1.19)

Risk-adjusted odds ratio of morbidity:

- Female (at BSA=1.8) v male (at BSA=2.0): 1.30 (1.24-1.36)
- Black v white (including patients with race other than white, black, Asian): 1.27 (1.18-1.36)
- Asian v white (including patients with race other than white, black, Asian): 1.16 (1.04 – 1.30)

For details of risk adjustment models, please see section 2b3 in the testing form.

The observed proportions of IMA use and perioperative medications were:

Observed proportions of IMA use:

- Female: 98.5% v male: 99.2%
- Black: 98.7% v non-black 99.1%:

Observed proportions of use of perioperative medications:

- Female: 92.6% v male: 92.5%
- Black: 93.2% v non-black: 92.5%

Note: Consistent with previous NQF reports, Non-North American hospitals are not included in models or percentages computations above. Results are virtually unchanged when these hospitals are included.This composite measure gauges the performance of the STS participant (typically a hospital, a hospital group, or a surgeon group). It is not a patient or operation level measure. Therefore at the composite score level, we do not provide data stratified by patient characteristics. Instead, we provide results stratified by participant characteristics.

Distribution of isolated CABG composite measures by regions, fall 2017 harvest, July 2016 - June 2017

Stat	Midwest	Northeast	South	West	Other*
# Participant	267	126	359	185	8
# Operations	33448	24800	63107	22601	1859
Mean	0.967	0.970	0.966	0.966	0.957
STD	0.00932	0.00848	0.0101	0.00958	0.0073
IQR	0.0109	0.0114	0.0119	0.0128	0.012
Percentiles					
0%	0.919	0.945	0.931	0.937	0.945
10%	0.957	0.958	0.951	0.953	0.948
20%	0.961	0.963	0.958	0.958	0.950
30%	0.964	0.967	0.962	0.961	0.952
40%	0.966	0.970	0.964	0.964	0.955

50%	0.969	0.972	0.966	0.967	0.958
60%	0.971	0.974	0.969	0.969	0.960
70%	0.973	0.976	0.971	0.972	0.962
80%	0.975	0.977	0.974	0.974	0.964
90%	0.978	0.979	0.978	0.977	0.964
100%	0.983	0.984	0.984	0.985	0.965

*Non-North American/Canadian

Distribution of isolated CABG composite measures by regions, fall 2016 harvest, July 2015- June 2016

Stat	Midwest	Northeast	South	West	Other*
# Participant	261	112	327	178	4
# Operations	32530	20875	55513	20427	627
Mean	0.968	0.971	0.966	0.965	0.960
STD	0.00911	0.00734	0.0109	0.0114	0.00938
IQR	0.0128	0.00887	0.0141	0.0144	0.00922
Percentiles					
0%	0.928	0.943	0.927	0.917	0.946
10%	0.956	0.963	0.952	0.952	0.950
20%	0.961	0.965	0.959	0.957	0.954
30%	0.964	0.968	0.962	0.961	0.958
40%	0.967	0.969	0.965	0.965	0.961
50%	0.969	0.972	0.968	0.968	0.962
60%	0.971	0.974	0.971	0.970	0.964
70%	0.974	0.975	0.973	0.972	0.965
80%	0.975	0.978	0.976	0.974	0.966
90%	0.978	0.980	0.979	0.977	0.966
100%	0.984	0.986	0.985	0.985	0.967

*Non-North American/Canadian

At the individual domain level, the risk-adjusted odds ratio associated with sex and race were:

Risk-adjusted odds ratio of mortality:

- Female (at BSA=1.8) v male (at BSA=2.0): 1.59 (95% confidence interval: 1.45-1.74)
- Black v white (including patients with race other than white, black, Asian): 1.17 (1.03 – 1.32)
- Asian v white (including patients with race other than white, black, Asian): 0.97 (0.80 – 1.19)

Risk-adjusted odds ratio of morbidity:

- Female (at BSA=1.8) v male (at BSA=2.0): 1.30 (1.24-1.36)
- Black v white (including patients with race other than white, black, Asian): 1.27 (1.18-1.36)
- Asian v white (including patients with race other than white, black, Asian): 1.16 (1.04 – 1.30)

For details of risk adjustment models, please see section 2b4.

The observed proportions of IMA use and perioperative medications were:

Observed proportions of IMA use:

- Female: 98.5% v male: 99.2%
- Black: 98.7% v non-black 99.1%:

Observed proportions of use of perioperative medications:

- Female: 92.6% v male: 92.5%
- Black: 93.2% v non-black: 92.5%

Note: Consistent with previous NQF reports, Non-North American hospitals are not included in models or percentages computations above. Results are virtually unchanged when these hospitals are included.This composite measure gauges the performance of the STS participant (typically a hospital, a hospital group, or a surgeon group). It is not a patient or operation level measure. Therefore at the composite score level, we do not provide data stratified by patient characteristics. Instead, we provide results stratified by participant characteristics.

Distribution of isolated CABG composite measures by regions, fall 2017 harvest, July 2016 - June 2017

Stat	Midwest	Northeast	South	West	Other*
# Participant	267	126	359	185	8
# Operations	33448	24800	63107	22601	1859
Mean	0.967	0.970	0.966	0.966	0.957
STD	0.00932	0.00848	0.0101	0.00958	0.0073
IQR	0.0109	0.0114	0.0119	0.0128	0.012
Percentiles					
0%	0.919	0.945	0.931	0.937	0.945
10%	0.957	0.958	0.951	0.953	0.948
20%	0.961	0.963	0.958	0.958	0.950
30%	0.964	0.967	0.962	0.961	0.952
40%	0.966	0.970	0.964	0.964	0.955
50%	0.969	0.972	0.966	0.967	0.958
60%	0.971	0.974	0.969	0.969	0.960
70%	0.973	0.976	0.971	0.972	0.962
80%	0.975	0.977	0.974	0.974	0.964
90%	0.978	0.979	0.978	0.977	0.964
100%	0.983	0.984	0.984	0.985	0.965

*Non-North American/Canadian

Distribution of isolated CABG composite measures by regions, fall 2016 harvest, July 2015- June 2016

Stat	Midwest	Northeast	South	West	Other*
# Participant	261	112	327	178	4
# Operations	32530	20875	55513	20427	627
Mean	0.968	0.971	0.966	0.965	0.960
STD	0.00911	0.00734	0.0109	0.0114	0.00938
IQR	0.0128	0.00887	0.0141	0.0144	0.00922
Percentiles					
0%	0.928	0.943	0.927	0.917	0.946

10%	0.956	0.963	0.952	0.952	0.950
20%	0.961	0.965	0.959	0.957	0.954
30%	0.964	0.968	0.962	0.961	0.958
40%	0.967	0.969	0.965	0.965	0.961
50%	0.969	0.972	0.968	0.968	0.962
60%	0.971	0.974	0.971	0.970	0.964
70%	0.974	0.975	0.973	0.972	0.965
80%	0.975	0.978	0.976	0.974	0.966
90%	0.978	0.980	0.979	0.977	0.966
100%	0.984	0.986	0.985	0.985	0.967

*Non-North American/Canadian

At the individual domain level, the risk-adjusted odds ratio associated with sex and race were:

Risk-adjusted odds ratio of mortality:

- Female (at BSA=1.8) v male (at BSA=2.0): 1.59 (95% confidence interval: 1.45-1.74)
- Black v white (including patients with race other than white, black, Asian): 1.17 (1.03 – 1.32)
- Asian v white (including patients with race other than white, black, Asian): 0.97 (0.80 – 1.19)

Risk-adjusted odds ratio of morbidity:

- Female (at BSA=1.8) v male (at BSA=2.0): 1.30 (1.24-1.36)
- Black v white (including patients with race other than white, black, Asian): 1.27 (1.18-1.36)
- Asian v white (including patients with race other than white, black, Asian): 1.16 (1.04 – 1.30)

For details of risk adjustment models, please see section 2b3 in the testing form.

The observed proportions of IMA use and perioperative medications were:

Observed proportions of IMA use:

- Female: 98.5% v male: 99.2%
- Black: 98.7% v non-black 99.1%:

Observed proportions of use of perioperative medications:

- Female: 92.6% v male: 92.5%
- Black: 93.2% v non-black: 92.5%

Note: Consistent with previous NQF reports, Non-North American hospitals are not included in models or percentages computations above. Results are virtually unchanged when these hospitals are included.

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.

Both options in 1c.1 describe the measure construction of the STS CABG Composite Score: Two or more individual performance measure scores (i.e. the 4 domains of this measure) combined into one score, and all-or-none measures (Domain 4). Additionally, Domain 2 has an "any-or-none" construction.

The STS CABG Composite Score measures surgical performance based on a combination of 11 NQF-endorsed process and outcomes measures. An NQF-endorsed structural measure, database participation, is included de facto as only STS Adult Cardiac Surgery Database participants are eligible to receive composite scores. To assess overall quality, the 11 NQF-endorsed measures are grouped into four domains, as described below.

Domain 1 – Absence of Operative Mortality

Proportion of patients (risk-adjusted) who do not experience operative mortality. Operative mortality is defined as death during the same hospitalization as surgery or after discharge but within 30 days of the procedure

Domain 2 – Absence of Major Morbidity

Proportion of patients (risk-adjusted) who do not experience any major morbidity. Major morbidity is defined as having at least one of the following adverse outcomes: 1) reoperations for any cardiac reason, 2) renal failure, 3) deep sternal wound infection, 4) prolonged ventilation/intubation, 5) cerebrovascular accident/permanent stroke

Domain 3 – Use of Internal Mammary Artery (IMA)

Proportion of first-time CABG patients who receive at least one IMA graft. Note: Patients with prior CABG surgery or with documented contraindication for IMA use (subclavian stenosis, previous cardiac or thoracic surgery, previous mediastinal radiation, an emergent or salvage procedure or no LAD disease) are not included in the denominator.

Domain 4 – Use of All Evidence-based Perioperative Medications

Proportion of patients who receive all required perioperative medications. The required perioperative medications are: 1) preoperative beta blockade therapy; 2) discharge anti-platelet medication; 3) discharge beta blockade therapy; and 4) discharge anti-lipid medication. Note: Patients who died before discharge are not eligible to receive the 3 discharge medications. Patients with a documented contraindication for any of the 4 medications are not eligible to receive that medication. No partial credit is given for a patient who received some, but not all, of the medications for which he or she is eligible.

The STS composite measure combines four separate domain-specific scores. Different weights were assigned to the four domains to reflect their relative importance. Hence the composite score is a weighted average across four domains. 81% of the total weight was assigned to mortality, 10% to morbidity, 7% to IMA and 3% to medications.

Participants receive a score for each of the four domains, plus an overall composite score. The overall composite score is created as a weighted average of the four domain scores, as described above. In addition to receiving a numeric score, participants are assigned to rating categories designated by one to three stars.

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. This composite provides a more comprehensive measure of overall quality.

The rationale for the use of a composite measure is threefold:

1. Data reduction – The CABG composite score simplifies evaluation of a cardiac surgery program by distilling all available NQF-endorsed measures into a simple summary.

2. Scope expansion – The CABG composite score is highly condensed, which makes it possible to track a broader range of metrics than would otherwise be possible, making provider assessments more comprehensive.

3. Provider performance valuation – If multiple indicators are to be used for evaluation of a cardiac surgery program, a method of translating several variables into a single value is necessary. This composite measure enables this translation, while giving the appropriate weight to each of the many variables.

Reference:

Peterson ED, DeLong ER, Masoudi FA, et al. ACCF/AHA 2010 position statement on composite measures for healthcare performance assessment: report of the American College of Cardiology Foundation/American Heart Association Task Force on Performance Measures (writing committee to develop a position statement on composite measures). Circulation 2010;121:1780-91.

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

For two of the domains (mortality; IMA usage), the study endpoint corresponds to a single measure. For the other two domains (morbidity; medications), the study endpoint is defined in a manner that combines multiple measures. For these two domains, the study endpoint is a composite endpoint.

A participant's overall composite performance score is calculated as a weighted average of the domain-specific estimates described above. The weight that is applied to a given domain is inversely proportional to the standard deviation of the domain-specific scores (calculated across hospitals).

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, Safety : Medication

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/documents/ACSD_DataCollectionFormV2_9_Annotated.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)*

Attachment **Attachment:** ACSD_DataSpecificationsV2_9.pdf

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.

None

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).

Please see Appendix

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).

Please see Appendix

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

Please see Appendix

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).

Please see Appendix

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

Please see Appendix

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.)*

Please see Appendix

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 (Appendix) and attached articles.

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 2.8 went live on July 1, 2014; STS Adult Cardiac Surgery Database Version 2.9 went live on July 1, 2017.

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

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

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

Clinician : Group/Practice, Facility

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

Inpatient/Hospital

If other:

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

Please see section S.4-S.11 (Appendix)

2. Validity – See attached Measure Testing Submission Form

comp_testing_v3.0_-_0696_STS_CABG_CompScore_112619update.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)

Measure Number (*if previously endorsed*)**:** 0696
Composite Measure Title: STS CABG Composite Score
Date of Submission: 8/1/2019

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
☐ other: Click here to describe	☐ other: Click here to describe

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 2.73

1.3. What are the dates of the data used in testing? July 2013 – 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: Click here to describe	☐ other: Click here to describe

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

The most recent isolated CABG composite measure included 143,771 operations from 1024 STS participants with surgery dates between July 2013 and June 2014. "Isolated CABG" is defined per the STS procedure table. All isolated CABG operations were included. The composite measures were calculated and reported for all participants who had more than 10 eligible cases during the one-year time window. Below is the distribution of sample sizes of all included participants. Frequency of included participants by geographic region is summarized in 1b.2.

Distribution of participant sample sizes (denominator), and event rates

Stat	N (Denominator)	% Mortality	% Morbidity	% no IMA	%failure to use medications
Mean	140.4	2.2	12.4	1.6	9.67
STD	110.2	2.1	6.5	2.6	11.42
IQR	122.5	2.2	7.8	2.3	11.07
Percentiles					
0%	10.0	0.0	0.0	0.0	0.00
10%	37.0	0.0	5.6	0.0	0.08
20%	55.0	0.6	7.4	0.0	1.47
30%	74.0	1.1	8.8	0.0	2.56
40%	91.0	1.5	10.3	0.3	4.02
50%	111.5	1.9	11.3	0.9	5.82
60%	133.0	2.2	12.8	1.3	8.32
70%	168.0	2.7	14.6	1.9	11.20
80%	211.0	3.5	16.9	2.7	15.76
90%	276.8	4.9	20.4	4.2	23.41
100%	942.0	17.6	51.2	29.4	79.49

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*)

All eligible patients were included except those missing a few key variables (age, gender), which STS never imputes in its risk adjustment algorithms.

	Effects	Overall N=143771
Age (years)	Median (IQR)	66.0 (58.0, 73.0)
	Missing	0 (0.0%)
Sex	Male	107,464 (74.7%)
	Female	36,307 (25.3%)
Race - Asian	No	139,141 (96.8%)
	Yes	4,241 (2.9%)
	Missing	389 (0.3%)
Race - Black / African American	No	132,497 (92.2%)
	Yes	10,895 (7.6%)
	Missing	379 (0.3%)
Race - White	No	21,618 (15.0%)
	Yes	121,894 (84.8%)
	Missing	259 (0.2%)
Race - American Indian / Alaskan Native	No	142,310 (99.0%)
	Yes	1,062 (0.7%)
	Missing	399 (0.3%)
Race - Other	No	138,421 (96.3%)
	Yes	4,819 (3.4%)
	Missing	531 (0.4%)
Native Hawaiian / Pacific Islander	No	142,713 (99.3%)
	Yes	591 (0.4%)
	Missing	467 (0.3%)
Hispanic or Latino Ethnicity	No	134,289 (93.4%)
	Yes	8,907 (6.2%)
	Missing	575 (0.4%)
Body Surface Area (m)	<1.5	2,157 (1.5%)
	>=1.5 and <1.75	17,854 (12.4%)
	>=1.75 and <2	49,637 (34.5%)
	>=2	74,042 (51.5%)
	Missing	81 (0.1%)
Diabetes	No Diabetes	76,816 (53.4%)
	Diabetes - Noninsulin	41,766 (29.1%)
	Diabetes - Insulin	24,761 (17.2%)
	Diabetes - Other	195 (0.1%)
	Diabetes - Missing Treatment	126 (0.1%)
	Missing	107 (0.1%)
Hypertension	No	16,789 (11.7%)
	Yes	126,924 (88.3%)
	Missing	58 (0.0%)
Renal Function	Creatinine <1 mg/dL	68,483 (47.6%)
	Creatinine 1-1.5 mg/dL	58,718 (40.8%)
	Creatinine 1.5-2 mg/dL	8,791 (6.1%)
	Creatinine 2-2.5 mg/dL	1,873 (1.3%)
	Creatinine >2.5 mg/dL	1,492 (1.0%)

	Effects	Overall N=143771
	Dialysis	3,987 (2.8%)
	Missing	427 (0.3%)
Chronic Lung Disease (CLD)	None	108,729 (75.6%)
	Mild	19,734 (13.7%)
	Moderate	8,641 (6.0%)
	Severe	6,412 (4.5%)
	Missing	255 (0.2%)
Peripheral Vascular Disease (PVD)	No	123,082 (85.6%)
	Yes	20,538 (14.3%)
	Missing	151 (0.1%)
Cerebrovascular Disease (CVD)	No	122,867 (85.5%)
	Yes	20,786 (14.5%)
	Missing	118 (0.1%)
Cerebrovascular Accident (CVA)	No CVA	133,006 (92.5%)
	Remote CVA (> 2 weeks)	10,218 (7.1%)
	Recent CVA (< 2 weeks)	352 (0.2%)
	CVA - Missing Timing	50 (0.0%)
	Missing	145 (0.1%)
Endocarditis	No Endocarditis	143,489 (99.8%)
	Treated Endocarditis	81 (0.1%)
	Active Endocarditis	18 (0.0%)
	Endocarditis - Missing Type	7 (0.0%)
	Missing	176 (0.1%)
Acuity Status	Elective	54,803 (38.1%)
	Urgent	82,336 (57.3%)
	Emergent	6,338 (4.4%)
	Emergent Salvage	229 (0.2%)
	Missing	65 (0.0%)
Myocardial Infarction	No Prior MI	68,828 (47.9%)
	MI >21 days	25,819 (18.0%)
	MI 8-21 days	6,755 (4.7%)
	MI 1-7 days	36,197 (25.2%)
	MI 6-24 hrs	3,751 (2.6%)
	MI <= 6 hrs	2,201 (1.5%)
	MI - Missing Timing	106 (0.1%)
	Missing	114 (0.1%)
Cardiogenic Shock	No	141,027 (98.1%)
	Yes	2,619 (1.8%)
	Missing	125 (0.1%)
Preop IABP	No	132,769 (92.3%)
	Yes	10,885 (7.6%)
	Missing	117 (0.1%)
Congestive Heart Failure	No CHF	116,909 (81.3%)
	CHF NYHA-I	2,483 (1.7%)

	Effects	Overall N=143771
	CHF NYHA-II	7,510 (5.2%)
	CHF NYHA-III	9,639 (6.7%)
	CHF NYHA-IV	6,278 (4.4%)
	CHF Missing NYHA	801 (0.6%)
	Missing	151 (0.1%)
Number of Diseased Coronary Vessels	None	428 (0.3%)
	One	5,832 (4.1%)
	Two	28,145 (19.6%)
	Three	109,160 (75.9%)
	Missing	206 (0.1%)
Left Main Disease > 50%	No	96,475 (67.1%)
	Yes	46,896 (32.6%)
	Missing	400 (0.3%)
Ejection Fraction (%)	Median (IQR)	55.0 (45.0, 60.0)
	Missing	3,936 (2.7%)
Dyslipidemia	No	17,768 (12.4%)
	Yes	125,933 (87.6%)
	Missing	70 (0.0%)
Aortic Stenosis	No	137,408 (95.6%)
	Yes	3,841 (2.7%)
	Missing	2,522 (1.8%)
Mitral Stenosis	No	139,593 (97.1%)
	Yes	439 (0.3%)
	Missing	3,739 (2.6%)
Tricuspid Stenosis	No	139,189 (96.8%)
	Yes	191 (0.1%)
	Missing	4,391 (3.1%)
Pulmonic Stenosis	No	140,011 (97.4%)
	Yes	97 (0.1%)
	Missing	3,663 (2.5%)
Aortic Insufficiency	None	122,179 (85.0%)
	Trivial	9,412 (6.5%)
	Mild	8,266 (5.7%)
	Moderate	1,805 (1.3%)
	Severe	50 (0.0%)
	Missing	2,059 (1.4%)
Mitral Insufficiency	None	82,497 (57.4%)
	Trivial	24,533 (17.1%)
	Mild	26,538 (18.5%)
	Moderate	7,564 (5.3%)
	Severe	594 (0.4%)
	Missing	2,045 (1.4%)
Tricuspid Insufficiency	None	88,334 (61.4%)
	Trivial	28,096 (19.5%)

	Effects	Overall N=143771
	Mild	20,967 (14.6%)
	Moderate	3,572 (2.5%)
	Severe	291 (0.2%)
	Missing	2,511 (1.7%)
Pulmonic Insufficiency	None	121,506 (84.5%)
	Trivial	13,844 (9.6%)
	Mild	5,041 (3.5%)
	Moderate	443 (0.3%)
	Severe	13 (0.0%)
	Missing	2,924 (2.0%)

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.

We used the same dataset of isolated CABG operations from July 2013 to June 2014 for the entire report (Fall 2014 harvest). The two exceptions are:

1. For empirical validity testing, we used the subset of participants who participated in STS during the time windows of July 2013 – June 2014 and July 2012 – June 2013.
2. In the individual measure risk prediction model validation section, we cited data from the paper published in 2009.

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:S2-S22.

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.

Whether outcomes measures, and the public reporting and reimbursement programs based on them, should consider socioeconomic (SES) or sociodemographic (SDS) factors (e.g., race, ethnicity, education, income, payer [e.g., Medicare-Medicaid dual eligible status]) is a topic of intense health policy debate [1]. 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. They also note that readily available SES factors have often not demonstrated significant impact on outcomes. As part of an NQF pilot project, STS specifically studied dual eligible status in the STS readmission measure [2] and found minimal impact. Finally, even SES proponents agree that these factors make more sense intuitively for some outcomes (e.g., readmission) than others (hospital mortality, complications)—that is, they are context-specific.

In identifying a risk adjustment approach for this measure, and in keeping with the general approach taken for the new STS risk models for the Adult Cardiac Surgery Database [3], we chose to avoid the more philosophical and downstream health policy implications of SES adjustment and based our modeling decisions on empirical

findings and consideration of the model's primary intended purpose--to adjust for case mix. 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 Adult Cardiac risk models as it has been previously, but not as a SES factor (nor as a surrogate for such factors). Race has an empirical association with outcomes [4,5] and has the potential to confound the interpretation of a hospital's outcomes, although the underlying mechanism is unknown (e.g., genetic factors, differential effectiveness of certain medications, rates of certain associated diseases not accounted for in the risk models, and racial differences in vessel anatomy and suitability for bypass). This is similar to the well-known fact that female gender is associated with worse outcomes, and is included in our CABG models (e.g. their coronary arteries tend to be smaller and more challenging for anastomoses).

Given the above explanation for the continued inclusion of race in our Adult Cardiac risk models, the STS agreed in an earlier measure review cycle (Fall 2018) to provide measure results stratified by race for future measure submissions. Please refer to the race-specific disparities data provided for each of the domains of measure 0696 (Mortality, Morbidity, IMA, Perioperative Medications) under question 1b.4 (Importance tab) of the submission form, which we believe will suffice for this new requirement.

1. National Academies of Sciences E, and Medicine. Accounting for social risk factors in medicare payment. Washington, DC: The National Academies Press; 2017.

2. Shahian DM, He X, O'Brien SM et al. Development of a clinical registry-based 30-day readmission measure for coronary artery bypass grafting surgery. Circulation 2014;130(5):399-409.

3. Shahian DM, Jacobs JP, Badhwar V, 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 May;105(5):1411-1418.

4. Bridges CR, Edwards FH, Peterson ED, Coombs LP. The effect of race on coronary bypass operative mortality. JACC 2000;36(6):1870-1876.

5. Bridges CR. Cardiac surgery in African Americans. Ann Thorac Surg 2003;76:S1356–62.

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*)

We used signal-to-noise ratio to measure the reliability of the composite measure. The method is described below.

Method details:

Reliability is conventionally defined as the proportion of variation in a performance measure that is due to true between hospital differences (ie, 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.

Let θ_j denote the true unknown composite measure value for the j-th of J hospitals. Prior to estimating reliability, the numerical value of θ_j was estimated for each hospital by the definition of CABG composite measure. Estimation was done using Markov Chain Monte Carlo (MCMC) simulations and involved the following steps:

1. *First, for each j, we randomly generated a large number N of possible numerical values of θ_j by sampling from the Bayesian posterior probability distribution of θ_j. Let $\theta_j^{(i)}$ denote the i-th of these N randomly sampled numerical values for the j-th hospital.*
2. *Second, for each j, a Bayesian estimate $\hat{\theta}_j$ 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 estimated 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

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

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*)

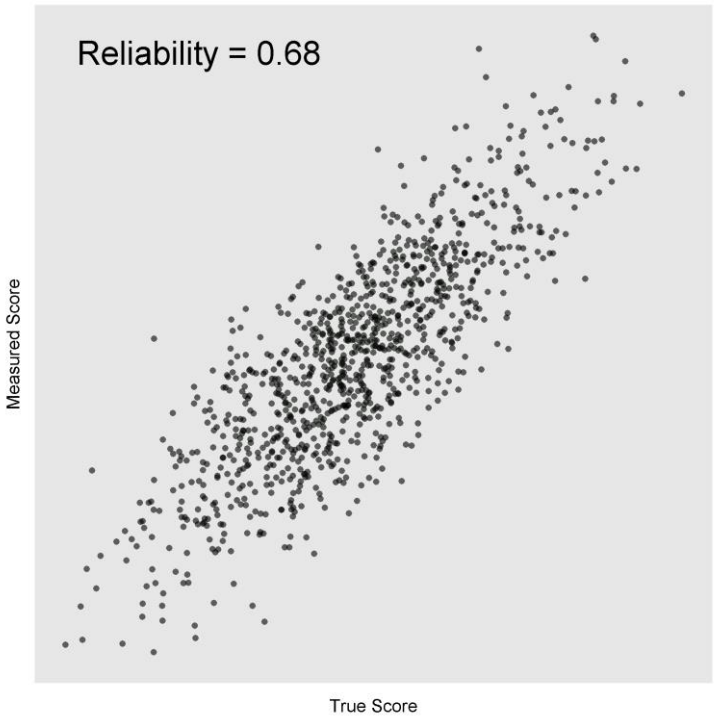
The posterior mean of reliability is 0.68. The posterior median, lower and upper boundaries of 95% credible intervals are 0.68, 0.63, and 0.73, respectively. The reliability is higher in participants with larger number of operations. The mean (95% credible interval) of reliability in participants with 50 or more operations (n = 863) is 0.71 (0.66, 0.76), and 0.72 (0.67, 0.77) in participants with 100 or more operations (n = 582).

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 result, we created a figure illustrating the accuracy of the measured scores when the true reliability = 0.68. Because the true score for the isolated CABG composite measure is unknown, we used simulated data with formula

$$MeasuredScore_i = TrueScore_i + e_i$$

Where i indicates observations, i=1 ... 1024, $TrueScore_i$ and e_i both follow normal distributions. The standard deviations of the normal distributions were chosen such that the measure has a reliability of 0.68.



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)*

Empirical validity testing

We tested the predictive validity of the composite measure. Predictive validity means that the results of this measure are predictive of future performance. We assessed the extent to which performance on the STS composite measure remains stable over time. In other words, does the composite score performed at one point in time accurately predict performance at some later time?

Face validity

Face validity implies that the measure is regarded as useful and valid by its intended users, including providers, consumers, payers, and regulators. This composite measure was developed with a panel of surgeon experts and statisticians. We have had near-universal acceptance of this measure by all stakeholders, and it has been used by a highly respected consumer publication, *Consumer Reports*, for nearly 5 years.

Content validity

Content validity means that the composite measure includes all of the essential dimensions of the underlying concept. STS believes that the four domains in the composite are broadly representative of the latent construct "isolated CABG quality".

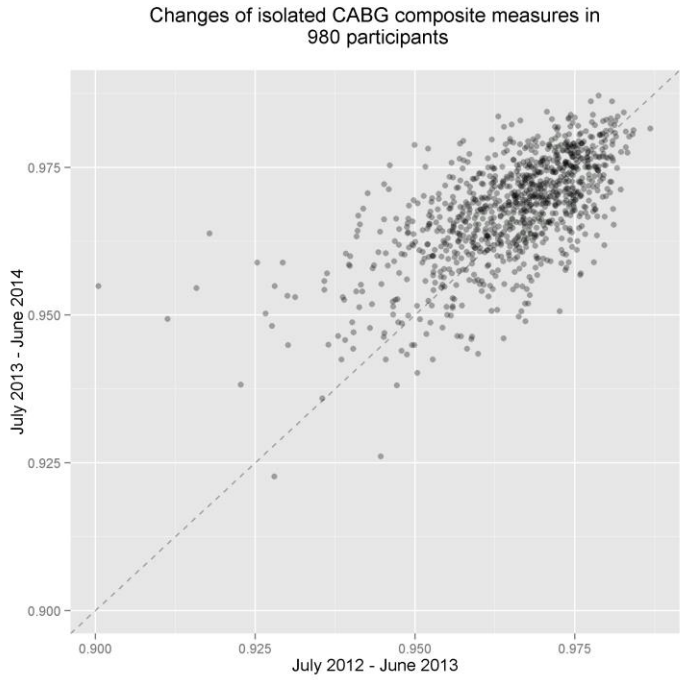
The tests on validity used the concept of performance outliers to be more formally introduced in 2b5: Participants were labeled as "better than average outliers" if it was at least 95% certain that the participant's true composite score was better than the overall STS average composite score. Participants were labeled as "worse than average outliers" if it was at least 95% certain the participant's true composite score was worse than the overall STS average composite score. Participants were labeled as better than average (3 stars), worse than average (1 star), and indistinguishable from the average (2 stars).

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

Empirical Validity

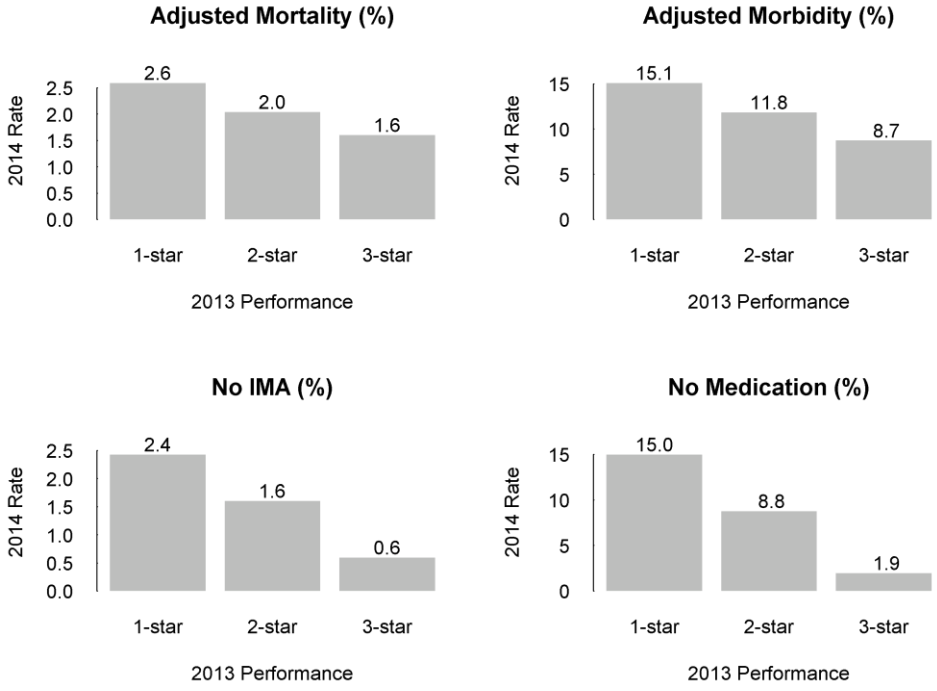
The analysis was restricted to a sample of 980 STS participants who participated and received a composite score in the last two non-overlapping harvest periods for which the isolated CABG composite score were reported (fall 2013, fall 2014).

Changes of scores between measures calculated with data in July 2013-June 2014 & July 2012-June 2013



The Pearson's correlation of the measure between the two time periods was 0.64; the Spearman's correlation was 0.63.

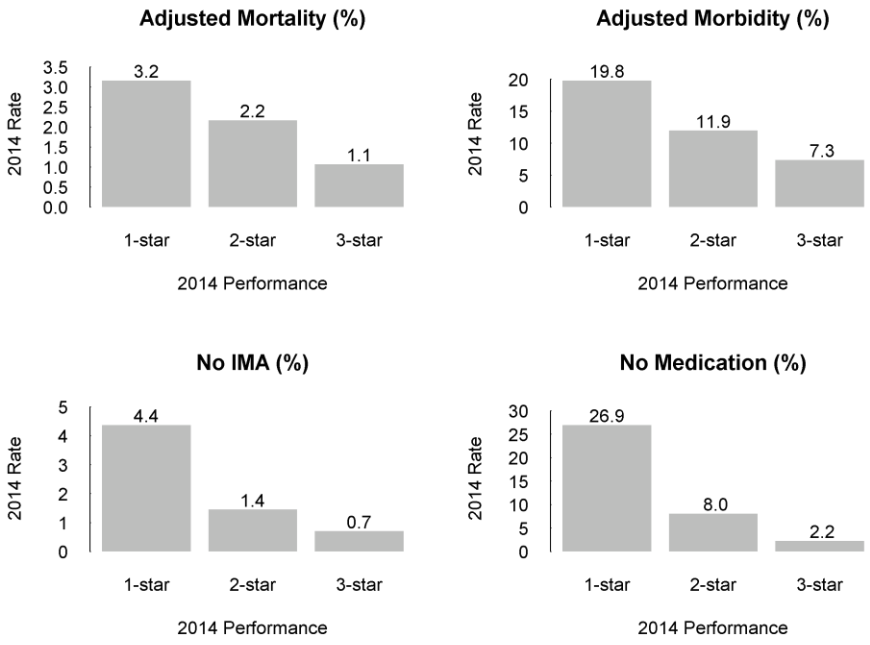
We also calculated the group-specific individual domain results from the later period (2014), by participants groups defined with their previous performance (2013).



Content Validity

STS participants deemed better by the composite scores have (on average) higher performance *during the same time window* on each individual domain of the composite measure. Thus, differences in performance were clinically meaningful as well as statistically significant. Compared to participants receiving 1 star, those

with 3 stars had better estimated performance for each individual domain of the composite score. This is illustrated in the figure below using data from fall 2014 harvest (July 2013 - June 2014). Compared to participants receiving 1 star, those with 3 stars had lower risk-adjusted mortality (1.1% vs. 3.2%) and lower risk-adjusted morbidity (7.3% vs. 19.8%).



2b1.4. What is your interpretation of the results in terms of demonstrating validity? (i.*e., what do the results mean and what are the norms for the test conducted?*)

The test results show that the composite measure encompasses all four domains as designed, and that the past measure can be used to predict future performance. Together with face value, they support the validity of the STS isolated CABG composite measure as a quality measure for isolated CABG.

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*)

For the IMA domain, we excluded patients who had a previous CABG prior to the current admission or if IMA was not used and one of the following reasons was provided:

- Subclavian stenosis
- Previous cardiac or thoracic surgery
- Previous mediastinal radiation
- Emergent or salvage procedure
- No LAD disease

For the medication domain, we applied different exclusions for different types of medications.

Medication	Exclusion
Preoperative beta blockade	Cases are removed from the denominator if preoperative beta blocker was contraindicated or if the clinical status of the patient was emergent or emergent salvage prior to entering the operating room
Beta blockade at discharge	Cases are removed from the denominator if there was an in-hospital mortality or if discharge beta blocker was contraindicated
Anti-platelet medication at discharge	Cases are removed from the denominator if there was an in-hospital mortality or if discharge aspirin was contraindicated
Anti-lipid treatment at discharge	Cases are removed from the denominator if there was an in-hospital mortality or if discharge anti-lipid treatment was contraindicated

To show the impact of the exclusions for each composite domain (and component measure) as described above, and how the measure would be distributed without it, we calculated and compared the distributions of the measure with and without the current exclusion criteria.

These individual domains/ component measures (with the exclusions) have all been endorsed by NQF.

0134 – Use of an internal mammary artery in CABG

0118 – Anti-lipid treatment at discharge

0116 – Anti-platelet medication at discharge

0117 – Beta blockade at discharge

0127 – Preoperative beta blockade

We include all operations in the mortality and morbidity domains.

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*)

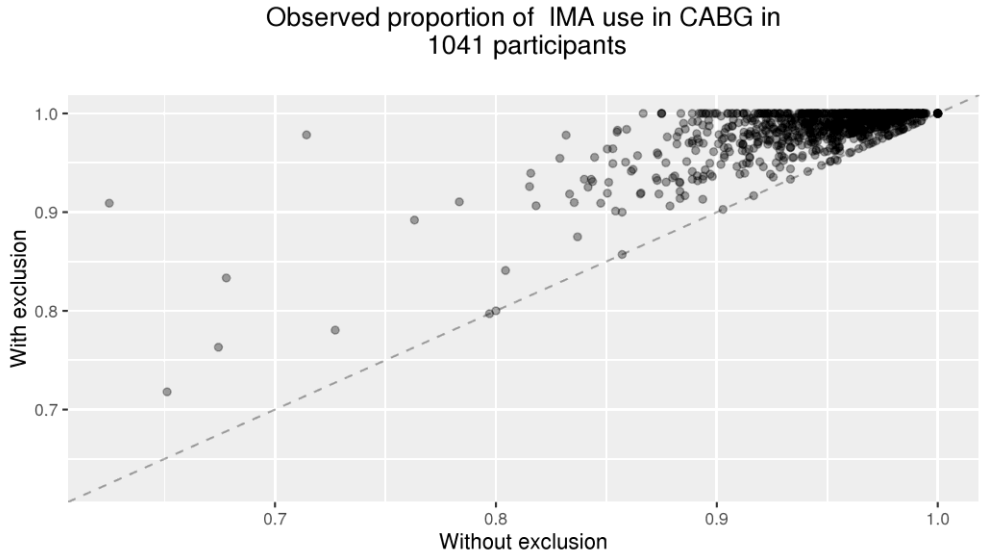
IMA domain

Distribution of participant-specific observed proportion of patients receiving the measure in October 2014 - September 2015 with and without the exclusion

Distribution	10/2014 – 09/2015 Observed proportion with exclusion	10/2014 – 09/2015 Observed proportion without exclusion
# Participant	1041	1041
# Operations	141347	147965
Mean	0.99	0.95
STD	0.026	0.043
IQR	0.019	0.046
0%	0.72	0.62
10%	0.96	0.90

20%	0.98	0.93
30%	0.98	0.94
40%	0.99	0.95
50%	0.99	0.96
60%	1.00	0.97
70%	1.00	0.98
80%	1.00	0.98
90%	1.00	1.00
100%	1.00	1.00
Low performance	76, 7.3%	111, 10.7%
Mid performance	944, 90.7%	810, 77.8%
High performance	21, 2.0%	120, 11.5%

Comparison of measure scores with and without the exclusion



The Spearman rank correlation of the measures with and without the exclusion is 0.60. The Pearson correlation is 0.73.

Medication domain

Anti-lipid treatment at discharge

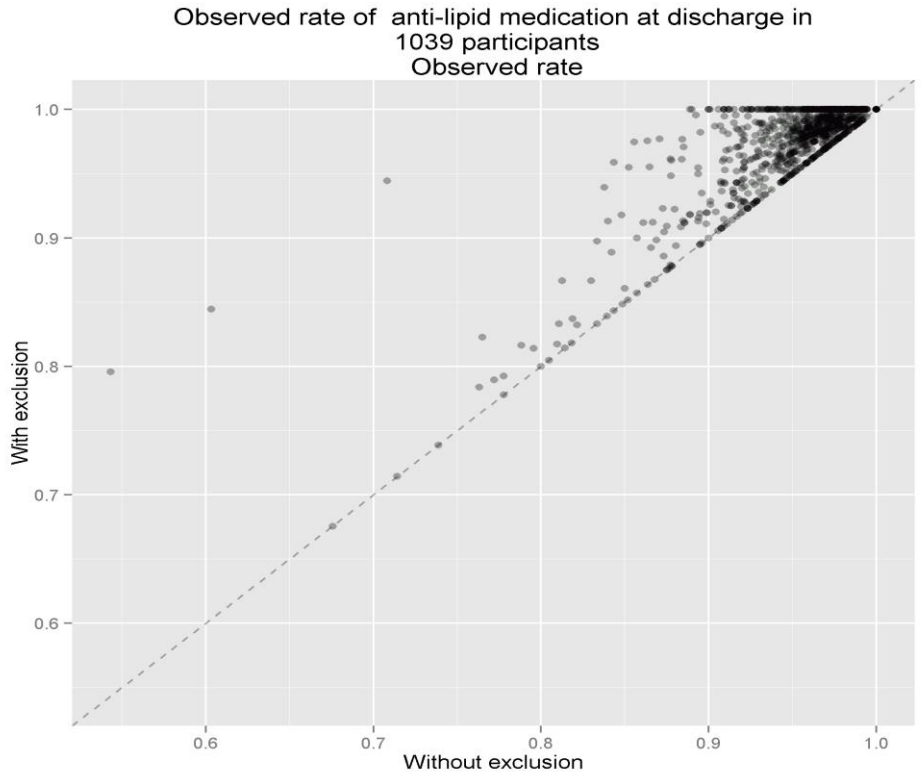
Distribution of participant-specific observed rates in July 2013 - June 2014 with and without the exclusion

Distribution	July 2013 - June 2014 Observed rate with exclusion	July 2013 - June 2014 Observed rate without exclusion
# Participant	1039	1039
# Operations	138587	142094
Mean	0.98	0.95
STD	0.038	0.045
IQR	0.032	0.04

0%	0.68	0.54
10%	0.93	0.91
20%	0.96	0.93
30%	0.97	0.95
40%	0.98	0.96
50%	0.99	0.96
60%	0.99	0.97
70%	1.00	0.98
80%	1.00	0.98
90%	1.00	0.99
100%	1.00	1.00
Midwest	300	300
Northeast	130	130
South	393	393
West	216	216
Low performance	98, 9.4%	92, 8.9%
Mid performance	840, 80.8%	837, 80.6%
High performance	101, 9.7%	110, 10.6%

The figure below shows the changes in participant specific observed rates with and without the exclusion criterion.

Comparison of measure scores with and without the exclusion



The Spearman rank correlation of the measures with and without the exclusion is 0.68. The Pearson correlation is 0.83.

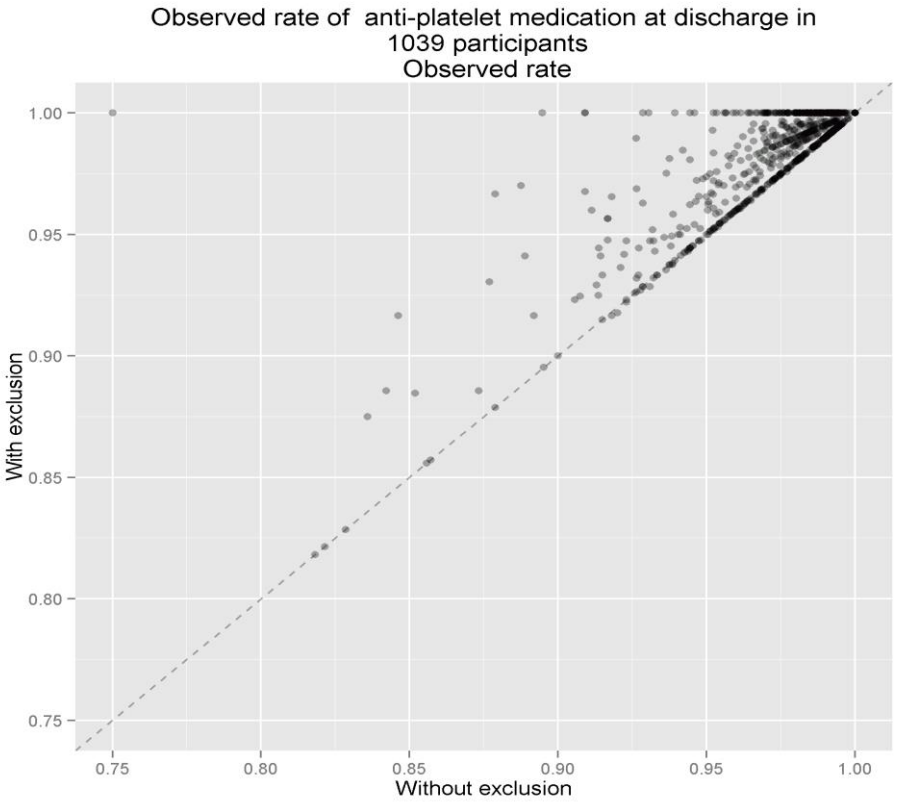
Anti-platelet medication at discharge

Distribution of participant-specific observed rates in July 2013 - June 2014 with and without the exclusion

Distribution	July 2013 - June 2014 Observed rate with exclusion	July 2013 - June 2014 Observed rate without exclusion
# Participant	1039	1039
# Operations	140024	141997
Mean	0.99	0.98
STD	0.022	0.027
IQR	0.017	0.028
0%	0.82	0.75
10%	0.96	0.95
20%	0.98	0.97
30%	0.99	0.98
40%	0.99	0.98
50%	1.00	0.99
60%	1.00	0.99
70%	1.00	1.00
80%	1.00	1.00
90%	1.00	1.00
100%	1.00	1.00
Midwest	300	300
Northeast	130	130
South	393	393
West	216	216
Low performance	80, 7.7%	80, 7.7%
Mid performance	939, 90.4%	914, 88.0%
High performance	20, 1.9%	45, 4.3%

The figure below shows the changes in participant specific observed rates with and without the exclusion criterion.

Comparison of measure scores with and without the exclusion



The Spearman rank correlation of the measures with and without the exclusion is 0.77. The Pearson correlation is 0.84.

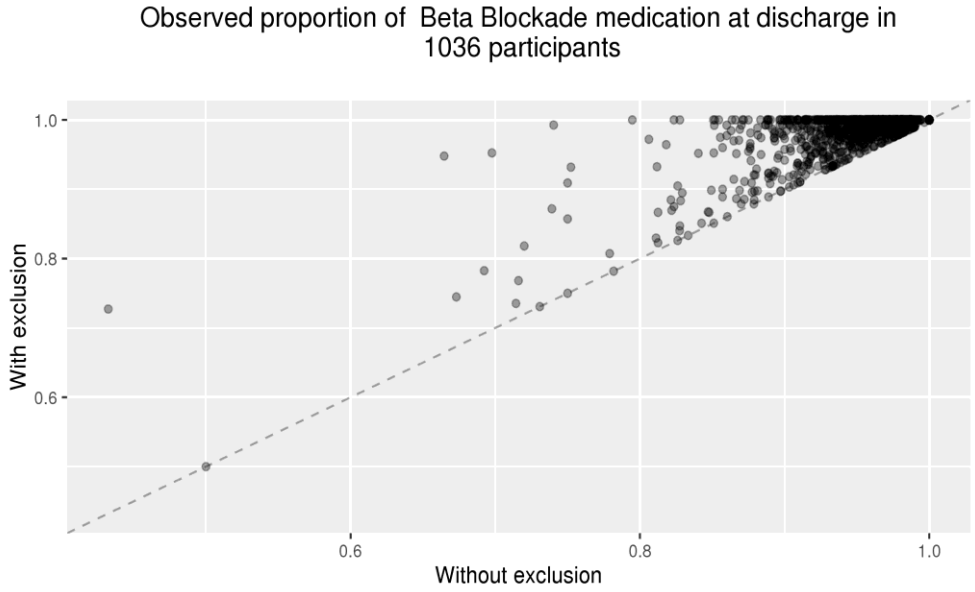
Beta blockade at discharge

Distribution of participant-specific observed proportion of patients receiving the measure in October 2014 - September 2015 with and without the exclusion

Distribution	10/2014 – 09/2015 Observed proportion with exclusion	10/2014 – 09/2015 Observed proportion without exclusion
# Participant	1036	1036
# Operations	139564	144880
Mean	0.98	0.94
STD	0.039	0.051
IQR	0.024	0.046
0%	0.50	0.43
10%	0.94	0.89
20%	0.97	0.92
30%	0.98	0.94

40%	0.99	0.95
50%	0.99	0.96
60%	1.00	0.96
70%	1.00	0.97
80%	1.00	0.98
90%	1.00	0.99
100%	1.00	1.00
Midwest	296	296
Northeast	136	136
South	389	389
West	215	215
Low performance	94, 9.1%	103, 9.9%
Mid performance	859, 82.9%	830, 80.1%
High performance	83, 8.0%	103, 9.9%

Comparison of measure scores with and without the exclusion



The Spearman rank correlation of the measures with and without the exclusion is 0.54. The Pearson correlation is 0.75.

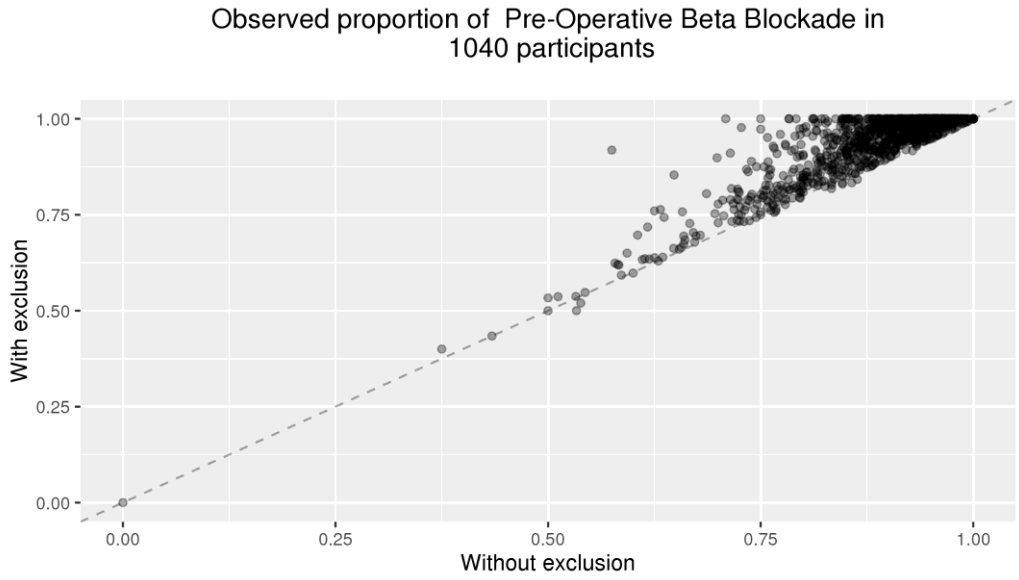
Preoperative beta blockade

Distribution of participant-specific observed proportion of patients receiving the measure in October 2014 - September 2015 with and without the exclusion

Distribution	10/2014 – 09/2015 Observed proportion with exclusion	10/2014 – 09/2015 Observed proportion without exclusion
# Participant	1041	1040

# Operations	134689	147699
Mean	0.93	0.88
STD	0.093	0.093
IQR	0.082	0.099
0%	0.00	0.00
10%	0.82	0.76
20%	0.89	0.82
30%	0.93	0.86
40%	0.95	0.89
50%	0.97	0.91
60%	0.98	0.93
70%	0.99	0.94
80%	1.00	0.95
90%	1.00	0.97
100%	1.00	1.00
Midwest	297	297
Northeast	136	136
South	392	391
West	216	216
Low performance	197, 18.9%	204, 19.6%
Mid performance	538, 51.7%	599, 57.6%
High performance	306, 29.4%	237, 22.8%

Comparison of measure scores with and without the exclusion



The Spearman rank correlation of the measures with and without the exclusion is 0.74. The Pearson correlation is 0.89.

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*)

For each composite domain (or component measure) to consistently quantify quality per its definition, it is necessary to exclude some cases, as specified in 2b2.1. The results of the exclusion analyses show that the exclusions have an impact on the results for many participants, and the results would be distorted without these appropriate exclusions.

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 48 risk factors**

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

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

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; the two domains that are not risk-adjusted or stratified are process measure domains.

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?

The mortality domain risk adjustment model has been endorsed by NQF. Please see measure forms for 0119 Risk-Adjusted Operative Mortality for CABG.

The morbidity domain risk adjustment model was developed in 2008 and published in 2009. The list of candidate risk predictors were picked by surgeon panel based on prior research and clinical expertise. Initial models were selected using a backwards approach with a significance criterion of 0.001 for removal. Age, body

surface area, surgery date in 6-month interval, ejection fraction, creatinine, sex and dialysis were preselected and forced into the models.

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:S2-S22.

The morbidity or mortality model is used to risk adjust the morbidity component of the composite measure. The composite mortality or morbidity in the aforementioned paper was defined in the same way as the morbidity component of the composite measure except that it also included mortality. Compared to morbidity, mortality is much rarer. The predictors of combined mortality and morbidity are essentially the same as the predicted risk of morbidity alone. At the participant level, raw morbidity rates and raw mortality or morbidity rates have very high correlation (Pearson=0.978, Spearman=0.972.) Because of this similarity, instead of devising a new model, we used a recalibrated version of the published and endorsed model for our any-or-none morbidity component in the composite measure.

Please also see 1.8 above regarding inclusion of race in the STS Adult Cardiac risk models.

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?

Effect	Estimated Odds Ratio of Any Adverse Event
Age 60 vs. 50	1.08 (1.05, 1.11)
Age 70 vs. 50	1.49 (1.44, 1.53)
Age 80 vs. 50	2.05 (1.98, 2.12)
BSA 1.6 vs. 2.0 among females	1.03 (1.01, 1.06)
BSA 1.6 vs. 2.0 among males	1.35 (1.30, 1.40)
BSA 1.8 vs. 2.0 among females	0.96 (0.94, 0.97)
BSA 1.8 vs. 2.0 among males	1.08 (1.07, 1.09)
BSA 2.2 vs. 2.0 among females	1.19 (1.16, 1.21)
BSA 2.2 vs. 2.0 among males	1.07 (1.06, 1.08)
Creatinine 1.5 vs. 1.0	1.76 (1.70, 1.82)
Creatinine 2.0 vs. 1.0	2.05 (1.98, 2.11)
Creatinine 2.5 vs. 1.0	2.39 (2.30, 2.48)
Dialysis vs. No Dialysis & Creat = 1.0	2.46 (2.33, 2.60)
EF - per 10 unit decrease	1.16 (1.15, 1.18)
Preop Afib	1.24 (1.21, 1.28)
CHF - not NYHA IV	1.27 (1.23, 1.31)
CHF - NYHA IV	1.48 (1.42, 1.54)

Chronic lung disease - mild	**1.23 (1.19, 1.27)**
Chronic lung disease - moderate	**1.42 (1.36, 1.47)**
Chronic lung disease - severe	**1.98 (1.90, 2.07)**
CVD with CVA	**1.32 (1.29, 1.36)**
CVD without CVA	**1.17 (1.14, 1.20)**
Diabetes - Insulin	**1.30 (1.27, 1.34)**
Diabetes - Noninsulin	**1.08 (1.06, 1.10)**
Diseased vessels (2 vs. 1 or 3 vs. 2)	**1.16 (1.14, 1.18)**
Preop IABP / Inotropes	**1.96 (1.86, 2.06)**
Shock	**2.10 (1.99, 2.23)**
Female vs. male (at BSA=1.8)	**1.18 (1.15, 1.21)**
Hypertension	**1.12 (1.10, 1.15)**
Immunosuppressive treatment	**1.20 (1.14, 1.26)**
Mitral Insufficiency - Moderate/Severe	**1.20 (1.15, 1.26)**
Tricuspid Insufficiency - Moderate/Severe	**1.24 (1.16, 1.33)**
PCI < 6 hours	**1.31 (1.23, 1.39)**
PVD	**1.25 (1.22, 1.28)**
Aortic stenosis	**1.16 (1.10, 1.22)**
Left main disease	**1.04 (1.02, 1.06)**
MI 1-21 days	**1.23 (1.20, 1.25)**
MI 6 to 24 hrs	**1.43 (1.37, 1.50)**
MI < 6 hrs	**1.44 (1.35, 1.53)**
Time Trend - Per 6 month harvest interval	**1.00 (1.00, 1.01)**
Race - Asian	**1.23 (1.15, 1.31)**
Race - Black	**1.31 (1.24, 1.38)**
Race - Hispanic	**1.12 (1.05, 1.19)**
Reop - 1 previous operation	**1.61 (1.50, 1.72)**
Reop - 2+ previous operations	**1.84 (1.65, 2.05)**
Status - Urgent	**1.18 (1.14, 1.22)**
Status - Emergent (no resuscitation)	**1.77 (1.64, 1.91)**
Status - Emergent w/ resuscitation or salvage	**3.65 (3.26, 4.09)**

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*)

In the paper referenced previously, the models were assessed for predictability by means of C-index and goodness-of-fit through calibration plot. Data were split into development and validation samples, and upon completion of model development, C-indices were estimated and calibration plots were created using the validation sample.

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*)**:**

C-statistic of the composite adverse event model: 0.725

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

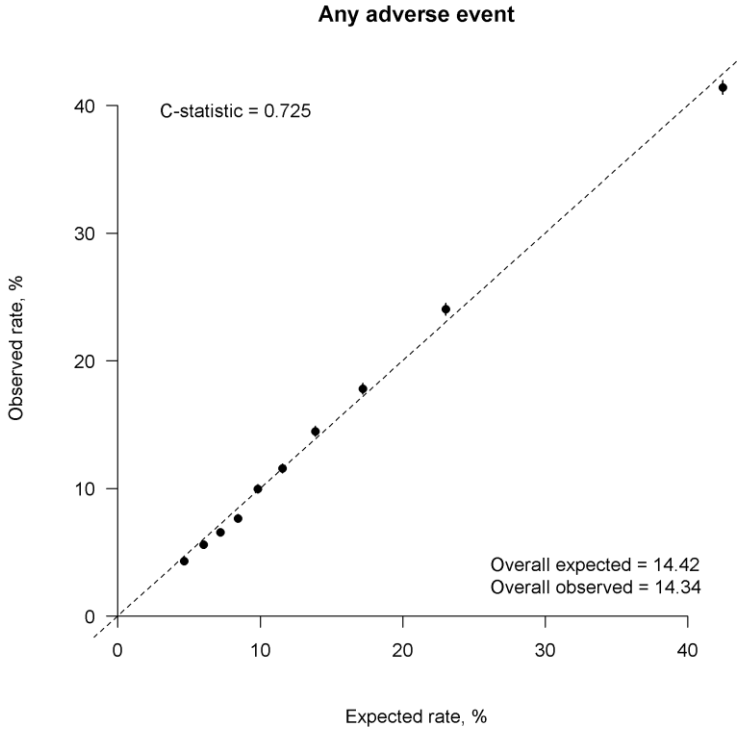
Calibration was assessed graphically by plotting observed versus expected event rates by decile of predicted risk overall and within several subgroups (2b4.8). The Hosmer-Lemeshow test was not used because it is known to be highly sensitive to sample size and is likely to be significant in studies with a very large sample size. We also calculated the absolute differences between observed and expected event rates in deciles defined by predicted risks. The largest such difference in any deciles is 1.0% for the composite adverse event model.

Observed and expected composite adverse event rates by risk deciles

	Decile 1	Decile 2	Decile 3	Decile 4	Decile 5	Decile 6	Decile 7	Decile 8	Decile 9	Decile 10
Observed composite adverse event rate, %	4.3	5.6	6.6	7.7	10.0	11.6	14.5	17.8	24.0	41.4
Expected composite adverse event rate, %	4.7	6.0	7.2	8.4	9.8	11.6	13.9	17.2	23.0	42.4

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

Calibration plot for mortality and morbidity composite endpoint in the validation sample



2b3.9. Results of Risk Stratification Analysis:

N/A

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

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

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

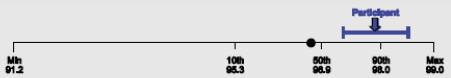
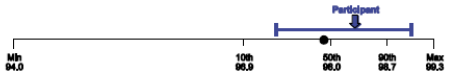
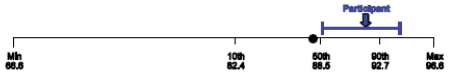
2b4. IDENTIFICATION OF STATISTICALLY SIGNIFICANT & MEANINGFUL DIFFERENCES IN PERFORMANCE

Note: *Applies to the composite performance measure.*

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

The degree of uncertainty surrounding an STS participant's composite measure estimate is indicated by calculating 98% Bayesian credible intervals (CI's) which are similar to conventional confidence intervals. Point estimates and CI's for an individual STS participant are reported along with a comparison to various

benchmarks based on the national sample. Benchmarks include the overall average STS composite score and several percentiles (minimum, 10th, 25th, 75th, 90th, maximum). A sample of the current STS reporting format is provided below. In addition, the composite measure result is converted into a star rating of 1 to 3 stars. An STS participant receives 2 stars if the Bayesian credible interval surrounding their composite score overlaps the overall STS average. This rating implies that the STS participant's performance was not statistically different from the overall STS national average. If the Bayesian CI falls entirely above the STS national average, the participant receives 3 stars (better than average performance). If the Bayesian CI falls entirely below the STS national average, the participant receives 1 star (worse than average performance).

Quality Domain	Participant Score (98% CI)	STS Mean Participant Score	Participant Rating[1]	Distribution of Participant Scores • = STS Mean
Jul 2013 - Jun 2014 Overall	97.9% (97.3 , 98.5)	96.7%	★★★	
Jul 2013 - Jun 2014 Absence of Mortality	98.3% (97.3 , 99.0)	97.9%	★★	
Jul 2013 - Jun 2014 Absence of Morbidity[2]	91.7% (88.6 , 94.1)	87.9%	★★★	
Jul 2013 - Jun 2014 Use of IMA[2]	99.4% (98.4 , 99.9)	98.4%	★★	
Jul 2013 - Jun 2014 Medications[2]	95.3% (93.0 , 97.1)	90.5%	★★★	

[1]* = Participant performance is significantly lower than the STS mean based on 99% Bayesian probability
* * = Participant performance is not significantly different than the STS mean based on 99% Bayesian probability
* * * = Participant performance is significantly higher than the STS mean based on 99% Bayesian probability
[2]Please refer to Report Overview - STS Composite Quality Rating and NQF-Endorsed Measures for full details

Quality Domain	Eligible Procedures	Detail	Count	Percent of Morbidity/Failure[1]
Jul 2013 - Jun 2014 Absence of Mortality	445	Mortality	5	
Jul 2013 - Jun 2014 Absence of Morbidity[2]	445	Any Morbidity	29	
		Reoperation only[3]	2	6.9 %
		Renal Failure only[4]	1	3.4 %
		Deep Sternal Infection/Mediastinitis only	0	0.0 %
		Prolonged Ventilation only	18	62.1%
		Cerebrovascular Accident only	4	13.8%
		Multiple Morbidities	4	13.8%
Jul 2013 - Jun 2014 Use of IMA[5]	435	IMA Failures	1	
Jul 2013 - Jun 2014 Medications[6]	445	Failed to Prescribe all eligible NQF-Endorsed Medications	24	
		Only Failed to Prescribe Preoperative Beta Blockade	16	66.7%
		Only Failed to Prescribe Discharge Beta Blockade[7]	3	12.5%
		Only Failed to Prescribe Discharge Anti-Lipids[7]	4	16.7%
		Only Failed to Prescribe Discharge Anti-Platelets[8]	0	0.0 %
		Failed to Prescribe Multiple Medications	1	4.2 %

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)

As shown in the table below, around 80% of participants have received 2 stars, and the remaining participants have received either 1 or 3 stars.

Star ratings in the last four harvests

Star Rating	07/2013-06/2014	01/2013-12/2013	07/2012-06/2013	01/2012-12/2012
1	60, 5.9%	98, 9.6%	97, 9.6%	91, 9.0%
2	864, 84.4%	770, 75.7%	782, 77.3%	770, 76.5%
3	100, 9.8%	149, 14.7%	132, 13.1%	146, 14.5%

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. We know with a 98% certainty that a 3-star participant have truly better performance, and a 1-star participant have truly worse performance. Better and worse are relative to the overall STS average. The identified differences in performance are both statistically significant and clinically meaningful. The surgeon panel and users are satisfied with the amount of outliers the measure detects.

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 nonresponders) and how the specified handling of missing data minimizes bias (*describe the steps—do not just name a method; what statistical analysis was used*)

Missing data on mortality and morbidity were extremely rare (~1/1000). We calculated the participant proportions of missing data on the IMA and medications and summarized the distributions. In the measure implementation, we did not include centers with more than 5% missing data, and imputed the missing data in the remaining centers to "no" (the undesirable value). To show if the exclusion of the centers due to missing data could introduce bias to the measure, we recalculated the measure including those centers and compared the results with existing ones.

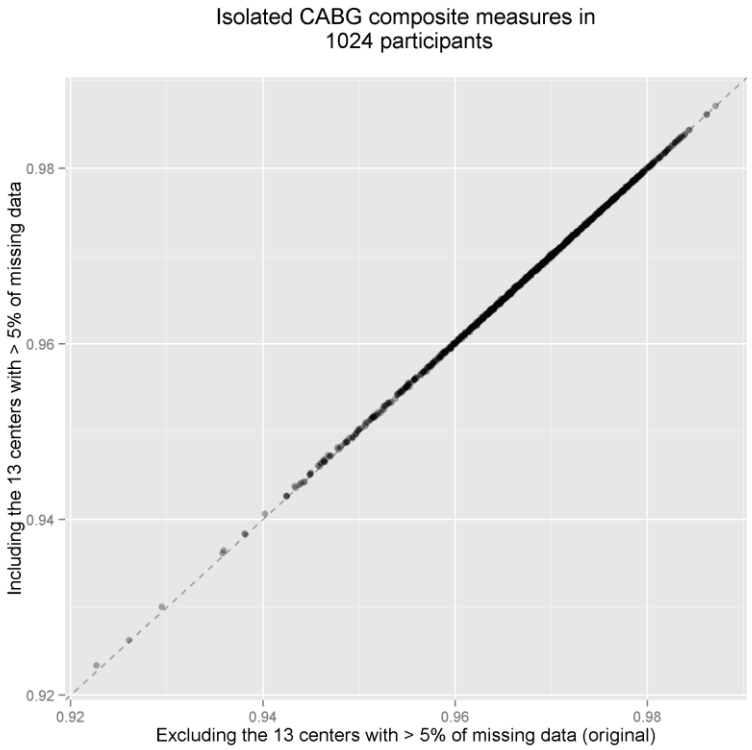
2b6.2. What is the overall frequency of missing data, the distribution of missing data across providers, and the results from testing related to missing data? (*e.g., results of sensitivity analysis of the effect of various rules for missing data/nonresponse; if no empirical sensitivity analysis, identify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

Distributions of missing data in IMA and medication domains across 1037 STS participants (July 2013-June 2014)

Percentiles	Use of IMA	Preoperative beta blockade	Beta blockade at discharge	Anti-platelet medication at discharge	Anti-lipid treatment at discharge
25th	0	0	0	0	0
50th (median)	0	0	0	0	0
75th	0	0	0	0	0
90th	0	0	0	0	0
99th	3.0%	1.5%	3.7%	3.7%	3.7%

Among the 1037 centers, 13 were excluded for due to high rate of missing data (> 5%) in least one field. On average, they were small centers and accounted for only 861 operations. In the remaining 1024 centers, we recalculated the composite measure including the 13 centers and the results were virtually identical.

Comparison of composite scores with and without the centers with > 5% missing in at least one field



The Pearson's correlation of the measure between the two was 0.9999, the Spearman's correlation was 0.9999, suggesting that the impact of missing data was minimal.

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

The rates of missing data in the STS Adult Cardiac Database were very low. We demonstrated that the exclusion of the centers with more than 5% of missing data on any of the domains does not have an impact on the measures of the other centers. We therefore concluded that systematic missing data did not lead to bias in our measure.

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 comprehensive score. We

also calculated the coefficients between the individual components of morbidity and the major morbidity domain score, as well as the coefficients between the individual medication failure rates and the perioperative medication domain score.

Details on methods for individual components of morbidity:

- Rates for each individual component are computed as Px=Number Cases / Total Patients for each Participant.
- Major morbidity composite rate: MM=Number of Cases with at least 1 of the 5 individual components present / Total Number of Patients.
- We compute correlation coefficients (Pearson and Spearman) between each individual component of the major morbidity composite and major morbidity rate.

Summary for each Major Complication Rate and Major Morbidity Domain Rate

CABG Population						
Major Morbidities	**N**	**Mean**	**Std Dev**	**Median**	**Minimum**	**Maximum**
Prolonged Ventilation	945	0.08196	0.05273	0.07308	0	0.34783
DSWI	945	0.00329	0.00678	0	0	0.04891
Permanent Stroke	945	0.01432	0.01432	0.01149	0	0.09677
Reoperation	945	0.02334	0.02051	0.01961	0	0.17778
New Case RF	945	0.02275	0.02171	0.01818	0	0.20000
Major Morbidity	945	0.11503	0.05780	0.10667	0	0.46154

Details on methods for individual medication failure rates:

- Rates for each individual medication are computed as FAILURE to receive medication: Px=Number Failure Cases / Total Patients for each Participant.
- Medications composite rate: MedFailure=Number of Cases FAILING to receive at least one medication / Total Number of Patients. This is an ALL or NONE measure. If at least one med was not received, then it is a FAILURE.
- We compute correlation coefficients (Pearson and Spearman) between each individual component of the medications composite and perioperative medication rate.

Summary for each Medication Failure Rate and Perioperative Medications Domain Rate

CABG Population **FAILURE to RECEIVE MEDs, mean proportion among participants**						
Meds / Composite	**N**	**Mean**	**Std Dev**	**Median**	**Minimum**	**Maximum**
Preoperative Beta-Blockers	945	0.05112	0.08022	0.01840	0	0.69231
Discharge Anti-Platelets	945	0.00945	0.01631	0	0	0.12500
Discharge Beta-Blockers	945	0.01583	0.03081	0.00441	0	0.30769
Discharge Anti-Lipids	945	0.02205	0.03819	0.01099	0	0.53846
Failure to Rececive ALL MEDs	**945**	**0.08449**	**0.10646**	**0.04706**	**0**	**0.84615**

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*)

The Pearson correlation coefficients were 0.48 between mortality and overall composite measure, 0.80 between morbidity domain score and overall score, 0.38 between IMA domain score and overall score, and 0.60 between medication domain score and overall score.

Correlation Coefficients between individual complications and major morbidity domain

CABG Population Correlation with Major Morbidity Domain Rate		
Major Morbidity Components	**Pearson Correlation Coefficient**	**Spearman Correlation Coefficient**
Prolonged Ventilation	0.90839	0.89068
DSWI	0.16573	0.10771
Permanent Stroke	0.34457	0.36699
Reoperation	0.44122	0.42296
New Case RF	0.45281	0.44247

Correlation Coefficients between individual medication failure rates and medications domain

CABG Population Correlation with Perioperative Medication Domain Rate		
Perioperative Medication Components	**Pearson Correlation Coefficient**	**Spearman Correlation Coefficient**
Preoperative Beta-Blockers	0.93674	0.87620
Discharge Anti-Platelets	0.51185	0.49528
Discharge Beta-Blockers	0.74638	0.63619
Discharge Anti-Lipids	0.72380	0.70072

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 much of the variation in the overall comprehensive score, it does not dominate. The other three domains also contribute statistical information.

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

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

The STS composite measure combines four separate domain-specific scores. Different weights were assigned to four domains to reflect their relative importance. Hence the composite score is a weighted average across four domains. 81% of the total weight was assigned to mortality, 10% to morbidity, 7% to IMA and 3% to medications.

In the original method development, the numbers were set to be the reciprocal of the standard deviation of the four domains before reviewed by the expert panel. Alternative methods of rescaling the individual scores were investigated in O'Brien et al.

O'Brien SM, Shahian DM, DeLong ER, Normand SL et al. Quality measurement in adult cardiac surgery: part 2--Statistical considerations in composite measure scoring and provider rating. Ann Thorac Surg 2007;83(4):S13-26.

The weights were deemed as an appropriate reflection of their relative importance by the expert panel. We calculated the relative importance between for the four domains. Specifically, we calculated for 1% change in mortality, how many percent change in each of the remaining domains is 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*)

1% change in mortality has the same impact on the overall composite score as

- 8.1% change in morbidity
- 11.6% chance in IMA
- 27.0% change in medications

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

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*)

The purpose of the composite score is to incorporate morbidity, mortality, use of IMA, and use of perioperative medications into a more comprehensive quality measure for isolated CABG quality. The calculated numbers passed the review by the expert panel.

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).

As of November 2019, the STS Adult Cardiac Surgery Database has 1,066 participants in the U.S. and Canada, 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 Adult Cardiac Surgery Database 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 Adult Cardiac Surgery Database 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 3 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 direct costs to collect the data for this measure. Costs to develop and maintain the measure included volunteer cardiothoracic surgeon time, STS staff time, and Duke Clinical Research Institute statistician and project management time.

Other fees:

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

4. Usability and Use

Extent to which potential audiences (e.g., consumers, purchasers, providers, policy makers) are using or could use performance results for both accountability and performance improvement to achieve the goal of high-quality, efficient healthcare for individuals or populations.

4a. Accountability and Transparency

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

4.1. Current and Planned Use

NQF-endorsed measures are expected to be used in at least one accountability application within 3 years and publicly reported within 6 years of initial endorsement in addition to performance improvement.

Specific Plan for Use	Current Use (for current use provide URL)
	Public Reporting STS Public Reporting https://publicreporting.sts.org/ STS Public Reporting https://publicreporting.sts.org/ Quality Improvement (Internal to the specific organization) STS Adult Cardiac Surgery Database Participants http://www.sts.org/sts-national-database/database-managers/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

Voluntary public reporting – approximately 72% of STS Adult Cardiac Surgery Database participants are enrolled as of November 2019. The STS CABG Composite has been publicly reported for consenting participants since 2010.
Quality improvement - STS Adult Cardiac Surgery Database Participant Feedback Reports provide performance results for this measure to participants. (see details in 4a2.1.1 below)

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

N/A

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

N/A

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

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

As of November 2019, there are 1,066 active U.S. and Canadian participants in the STS Adult Cardiac Surgery Database (ACSD). A "participant" is a cardiothoracic surgeon or group of cardiothoracic surgeons who agree to submit case records for analysis and comparison with benchmarking data for quality improvement initiatives. At the option of the surgeon or surgical group, the ACSD participant can include a hospital and/or associated anesthesiologists. It is for this reason that we have indicated (on the Specifications tab, question #S.20) that this measure is specified/tested for both the "clinician: group/practice" and "facility" levels of analysis.

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

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

Please see response under 4a2.1.1

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

Describe how feedback was obtained.

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

Most recently, STS surgeon members have expressed interest in real-time, online data updates, which has led to the development of dashboard-type reporting on STS.org. Roll-out of the adult cardiac dashboard is underway in 2019.

Also, adult cardiac public reporting has been available since 2010 (http://publicreporting.sts.org/acsd), making star ratings for consenting participant groups available to participants as well as the public.

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

Please see response under 4a2.2.1

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

Voluntary participation in ACSD public reporting has continually increased over the years that the initiative has been available, from 38% of ACSD participants in 2014, to 49% in 2016, to approximately 72% as of November 2019. This trend suggests that feedback from ACSD participants and others who access the performance data available on STS.org is sufficiently positive to promote ever-increasing participation in public reporting.

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

N/A

Improvement

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

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

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

Please see table below displaying 2010-2018 star ratings for this measure, in percentages. (If table below does not display clearly, please see version in Appendix.)

The data demonstrate that the general trend since 2010 has been a decrease in the percentage of surgical programs with 1-star and 3-star ratings on the CABG Composite, 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.

Stars	2018	2017	2016	2015	2014	2013	2012	2011	2010
*	4.37	4.55	5.29	5.82	4.59	9.19	9.0	9.6	11.0
**	88.27	89.21	84.65	84.4	86.64	75.86	76.0	76.5	75.5
***	7.36	6.24	10.00	9.74	8.77	14.95	15.0	14.0	13.5

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 2019, as in each year since 2014. 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.

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)

0114 : Risk-Adjusted Postoperative Renal Failure

0115 : Risk-Adjusted Surgical Re-exploration

0116 : Anti-Platelet Medication at Discharge

0117 : Beta Blockade at Discharge

0118 : Anti-Lipid Treatment Discharge

0119 : Risk-Adjusted Operative Mortality for CABG

0120 : Risk-Adjusted Operative Mortality for Aortic Valve Replacement (AVR)

0121 : Risk-Adjusted Operative Mortality for Mitral Valve (MV) Replacement

0122 : Risk-Adjusted Operative Mortality for Mitral Valve (MV) Replacement + CABG Surgery

0123 : Risk-Adjusted Operative Mortality for Aortic Valve Replacement (AVR) + CABG Surgery

0127 : Preoperative Beta Blockade

0129 : Risk-Adjusted Postoperative Prolonged Intubation (Ventilation)

0130 : Risk-Adjusted Deep Sternal Wound Infection

0131 : Risk-Adjusted Stroke/Cerebrovascular Accident

0134 : Use of Internal Mammary Artery (IMA) in Coronary Artery Bypass Graft (CABG)

1501 : Risk-Adjusted Operative Mortality for Mitral Valve (MV) Repair

1502 : Risk-Adjusted Operative Mortality for Mitral Valve (MV) Repair + CABG Surgery

2514 : Risk-Adjusted Coronary Artery Bypass Graft (CABG) Readmission Rate

2683 : Risk-Adjusted Operative Mortality for Pediatric and Congenital Heart Surgery

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

The following additional related measures were not accessible with the search function in 5.1a. All of these measures are NQF endorsed.

2561 STS Aortic Valve Replacement (AVR) Composite Score

2563 STS Aortic Valve Replacement (AVR) + Coronary Artery Bypass Graft (CABG) Composite Score

3030 STS Individual Surgeon Composite Measure for Adult Cardiac Surgery

3031 STS Mitral Valve Repair/Replacement (MVRR) Composite Score

3032 STS MVRR + CABG Composite Score

3294 STS Lobectomy for Lung Cancer 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:** 0696_Appx2019-S4-S11-1b2-1b4-CABG_2007-09-18_Papers-4b1.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 National Databases 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; All Children's Hospital/John Hopkins University, Saint Petersburg, 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: 2007

Ad.3 Month and Year of most recent revision: 01, 2015

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

Ad.5 When is the next scheduled review/update for this measure? 01, 2020

Ad.6 Copyright statement:

Ad.7 Disclaimers:

Ad.8 Additional Information/Comments: For completeness, several documents that were provided to the NQF Scientific Methods Panel for their October 2019 review of measure 0696 are included in the Appendix:

- Template Developer Response to SMP Prelim. Analysis (pg. 46-47)
- STS 2018 Adult Cardiac Surgery Risk Models: Part 1, Part 2 (pg. 48-65)
- STS CABG Composite for NQF Partial Update 2018 (pg. 66)