0003-4975/$36.00 467
https://doi.org/10.1016/j.athoracsur.2021.06.084

A REPORT OF THE STS QUALITY MEASUREMENT TASK FORCE AND THE STS WORKFORCE ON QUALITY

The STS Participant-Level, Multiprocedural Composite Measure for Adult Cardiac Surgery

David M. Shahian, MD, Vinay Badhwar, MD, Paul A. Kurlansky, MD, Michael E. Bowdish, MD, MS, Kevin W. Lobdell, MD, Anthony P. Furnary, MD, Vinod H. Thourani, MD, Jeffrey P. Jacobs, MD, Moritz C. Wyler von Ballmoos, MD, PhD, Karen M. Kim, MD, Christina Vassileva, MD, Mark S. Antman, DDS, MBA, Maria V. Grau-Sepulveda, MD MPH, and Sean M. O'Brien, PhD

Division of Cardiac Surgery, Department of Surgery, and Center for Quality and Safety, Massachusetts General Hospital, and Harvard Medical School, Boston, Massachusetts; Department of Cardiovascular and Thoracic Surgery, West Virginia University, Morgantown, West Virginia; College of Physicians and Surgeons, Columbia University, New York, New York; Departments of Surgery and Preventive Medicine, Keck School of Medicine of USC, University of Southern California, Los Angeles, California; Atrium Health, Cardiovascular and Thoracic Surgery, Charlotte, North Carolina; Starr-Wood Cardiothoracic Group, Portland, Oregon; Department of Cardiovascular Surgery, Marcus Valve Center, Piedmont Heart Institute, Atlanta, Georgia; Division of Thoracic and Cardiovascular Surgery, Department of Surgery, University of Florida, Gainesville, Florida; Houston Methodist DeBakey Heart and Vascular Center, Houston, Texas; Department of Cardiac Surgery, University of Michigan, Ann Arbor, Michigan; Integrity Medical Consulting, LLC, Shrewsbury, Massachusetts; The Society of Thoracic Surgeons, Chicago, Illinois; and Duke Clinical Research Institute, Duke University Medical Center, Durham, North Carolina

ABSTRACT

BACKGROUND Composite performance measures for the Society of Thoracic Surgeons (STS) Adult Cardiac Surgery Database participants (typically hospital departments or practice groups) are currently available only for individual procedures. To assess overall participant performance, STS has developed a composite metric encompassing the most common adult cardiac procedures.

METHODS Analyses included 1-year (July 1, 2018 to June 30, 2019) and 3-year (July 1, 2016 to June 30, 2019) time windows. Operations included isolated coronary artery bypass grafting (CABG), isolated aortic valve replacement (AVR), isolated mitral valve repair (MVr) or replacement (MVR), AVR + CABG, MVr or MVR + CABG, AVR + MVr or MVR, and AVR + (MVr or MVR) + CABG. The composite was estimated using Bayesian hierarchical models with risk-adjusted mortality and morbidity end points. Star ratings were based upon whether the 95% credible interval of a participant's score was entirely lower than (1 star), overlapping (2 star), or higher than (3 star) the STS average composite score.

RESULTS The North American procedural mix in the 3-year study cohort was as follows: 448 569 CABG, 72 067 AVR, 35 708 MVr, 29 953 MVR, 45 254 AVR + CABG, 12 247 MVr + CABG, 10 118 MVR + CABG, 3743 AVR + MVr, 6846 AVR + MVR, and 3765 AVR + (MVr or MVR) + CABG. Mortality and morbidity weightings were similar for 1- and 3-year analyses (76% and 24% [3-year]), as were composite score distributions (median, 94.7%; interquartile range, 93.6% to 95.6% [3-year]). The 3-year time frame was selected for operational use because of higher model reliability (0.81 [0.78-0.83]) and better outlier discrimination (26%, 3 star; 16%, 1 star). Risk-adjusted outcomes for 1-, 2-, and 3-star programs were 4.3%, 3.0%, and 1.8% mortality and 18.4%, 13.4%, and 9.7% morbidity, respectively.

CONCLUSIONS The STS participant-level, multiprocedural composite measure provides comprehensive, highly reliable, overall quality assessment of adult cardiac surgery practices.

(Ann Thorac Surg 2022;114:467-75)

The Supplemental Material can be viewed in the online version of this article [10.1016/j.athoracsur.2021.06.084] on http://www.annalsthoracicsurgery.org.

Accepted for publication Jun 28, 2021.
Presented at the Fifty-seventh Annual Meeting of The Society of Thoracic Surgeons, Virtual Meeting, Jan 29-31, 2021.
The Society of Thoracic Surgeons Executive Committee approved this document.
Address for correspondence: Dr Shahian, Division of Cardiac Surgery, Department of Surgery and Center for Quality and Safety, Massachusetts General Hospital, Bulfinch 282, 55 Fruit St, Boston, MA, 02114; email: dshahian@partners.org.

Ann Thorac Surg
2022;114:467-75

In the United States, ongoing efforts to assess and improve cardiac surgery performance have included the following: states (eg, New York, Pennsylvania, New Jersey, California, and Massachusetts) that led early public reporting initiatives; regional collaboratives (eg, the Northern New England Cardiovascular Disease Study Group, the Michigan Society of Thoracic and Cardiovascular Surgeons [MSTCVS] and MSTCVS Quality Collaborative, and the Virginia Cardiac Services Quality Initiative) that have shared best practices to promote quality improvement and cost reduction; the Department of Veterans Affairs, which centrally reviews data from all Veterans Affairs cardiac surgery programs to identify and remediate low-performing centers; and The Society of Thoracic Surgeons (STS), whose National Database generates comprehensive, risk-adjusted participant feedback reports to facilitate improvement activities and serves as the foundation for a robust voluntary public reporting program.[1-4]

For related articles, see pages 366 and 368

EVOLUTION OF STS ADULT CARDIAC SURGERY COMPOSITE MEASURES. Early quality assessment initiatives in cardiac surgery focused mainly on 1 procedure—coronary artery bypass grafting (CABG)—and 1 outcome, risk-adjusted mortality. Subsequently, given the progressive decline in CABG volumes and the desire to include a greater proportion of a typical practice's cases, risk-adjusted mortality metrics for other individual procedures were added, including valve replacements with or without concomitant CABG. Selected complications were sometimes included in feedback reports to providers, but rarely if ever were these publicly reported.

In 2007, recognizing the limited dimensionality (ie, mortality only) of existing cardiac surgery quality measures, the STS Quality Measurement Task Force (QMTF) developed the first of a series of STS participant-level (typically a hospital department or group practice) composite performance measures. The STS CABG composite measure[5,6] included 2 outcomes domains (risk-adjusted operative mortality and risk-adjusted, any-or-none morbidity [sternal infection, reoperation, renal failure, stroke, or prolonged ventilation]), as well as 2 process measure domains (use of at least 1 internal thoracic artery graft, and use of all 4 National Quality Forum-endorsed perioperative medications [preoperative β-blockers; discharge lipid-lowering agents, antiplatelet drugs, and β-blockers]), with domain-weighting determined empirically by the inverse of the standard deviations of each individual measure. Selection of the 5 major complications included in the any-or-none morbidity domain was made on the basis of their severity, as reflected by their potential to be associated with death or serious disability. For example, as part of a recent failure to rescue analysis, QMTF studied the mortality rates (across all STS adult cardiac index procedures) associated with various postoperative complications. Compared with patients who had no complications, who had an observed mortality rate of 0.9%, the mortality rate for patients who had only prolonged ventilation was 8.8%, 10-fold greater. The observed mortality rate for patients who had only a reoperation was 6.4%, 7-fold greater. Comparable mortality rates for patients experiencing only renal failure or stroke were 8.9% and 4.2%, respectively. Including patients who had prolonged ventilation together with at least 1 other complication, the associated mortality rate was 17.6%; similarly, reoperation combined with at least 1 other complication was associated with a mortality rate of 16.3%.

This composite measure provided a much more comprehensive assessment of performance and also facilitated identification of providers with better or worse than expected performance because of a larger number of end points. For example, only 1% of STS Adult Cardiac Surgery Database (ACSD) participants could be identified as outliers when using risk-adjusted operative mortality alone (with 98% credible intervals [CrIs]), compared with 23% outliers identified using the composite measure.

Subsequently, this composite approach has been expanded to additional, commonly performed adult cardiac procedures: isolated aortic valve replacement (AVR),[7] AVR + CABG,[8] mitral valve replacement (MVR) or repair (MVr),[9] and MVR or MVr + CABG.[10] These measures incorporate only risk-adjusted mortality and morbidity domains because nationally endorsed, widely accepted process measures were not available for many valve procedures.

STS INDIVIDUAL SURGEON, MULTIPROCEDURAL COMPOSITE MEASURE. All these STS metrics were at the STS-participant level, but there is increasing national interest in also assessing the performance of individual surgeons. This is methodologically challenging because the numbers of any specific procedure performed by individual surgeons are relatively small, even with multiple years of data. Accordingly, in 2015, the QMTF developed and published its first multiprocedural composite measure to address the surgeon sample size issue and, more broadly, to encompass a surgeon's overall practice.[11] This measure includes data from 5 procedures (representing, on average, 78% of a typical surgeon's practice), 2 outcomes domains, and 3 years of data. Because of the large number of end points, the reliability of this measure was the highest of any STS performance measure (0.81).

STS PARTICIPANT-LEVEL MULTIPROCEDURAL COMPOSITE MEASURE. In 2021, QMTF presents the next evolutionary step in adult cardiac surgery performance assessment—an STS participant-level, multiprocedural composite measure. There is increasing interest from patients, governmental and commercial payers, and regulators for

a comprehensive measure of a hospital or practice group's overall performance encompassing the most common adult cardiac procedures. To address this need, QMTF has developed a multiprocedural composite measure at the level of STS participants that uses the same general modeling strategy used in the individual surgeon, multiprocedural composite measure.[11] This new composite includes the same risk-adjusted mortality and morbidity outcomes domains and the same portfolio of common adult cardiac procedures: isolated CABG, isolated AVR, AVR + CABG, isolated MVR or MVr, MVR + CABG, and MVr + CABG. In addition, because of their increasing frequency, multiple valve procedures with or without CABG were also included.

As with the individual surgeon composite, the participant-level composite measure was designed to reflect the proportion of various case types performed by each participant. For example, those participants focusing on structural heart disease would have more of their overall score determined on the basis of valve procedures compared with a program focusing on ischemic heart disease.

MATERIAL AND METHODS

COHORT INCLUSION. Separate exploratory analyses were performed using 1-year and 3-year cohorts. The range of surgery dates was July 1, 2018 to June 30, 2019 in the 1-year cohort and July 1, 2016 to June 30, 2019 in the 3-year analysis. Only North American sites were included.

To assemble the study cohort, we first identified all operations that met inclusion criteria for the STS 2018 risk models[12,13] or the newly developed STS 2021 multiple valve with or without CABG risk models.[14] The starting population was 233 600 records from 1020 North American participants in the 1-year analysis and 715 333 from 1093 North American participants in the 3-year analysis. To be included in the 1-year analysis, participants had to have ≤2% missing or unknown operative mortality data in 2018 and 2019 and at least 10 eligible operations during that period. To be included in the 3-year analysis, participants were required to have missing or unknown operative mortality ≤5% in 2016 and ≤2% in all subsequent years and to have at least 10 eligible operations during those 3 years. The final study population was 220 081 records from 930 sites in the 1-year analysis and 668 270 records from 977 sites in the 3-year analysis.

MODEL OUTCOMES. The STS participant-level, multiprocedural composite score combines 2 outcome domains: risk-standardized mortality and risk-standardized major morbidity. Operative mortality is defined as death before hospital discharge or within 30 days of the operation. Major morbidity is defined as the occurrence of any 1 or more of the following major complications: prolonged ventilation, deep sternal infection, permanent stroke, renal failure, and reoperations for bleeding, coronary graft occlusion, prosthetic or native valve dysfunction, or other cardiac reasons, but not for other noncardiac reasons. This any-or-none morbidity domain is patient-centric because it identifies patients who achieve the optimal result of being discharged alive and without any of these 5 major complications.

STATISTICAL ANALYSIS. Detailed statistical methods are provided in the Supplemental Appendix. Briefly, participant-specific risk-adjusted mortality and morbidity rates were estimated in a Bayesian bivariate random-effects logistic regression model. Analyses were performed separately using the 1-year and 3-year study cohorts. For each outcome individually, the form of the model was a logistic regression with participant-specific random effects (intercepts). Random effects for mortality and morbidity were estimated jointly and were assumed to be correlated. Before fitting the bivariate random effects model, we first estimated risk scores predicting each outcome in a series of univariate logistic regression models omitting the random effect parameters. The goal of estimating risk scores was to reduce the number of covariates in the bivariate random effects model by summarizing the predictive information from many baseline covariates into a single number. The resulting risk scores were used as covariates in the bivariate random effects models.

Risk scores were estimated using the STS 2018 models[12,13] for patients undergoing CABG, AVR, MVR or MVr, AVR + CABG, and (MVR or MVr) + CABG and the STS 2021 multiple valve (with or without CABG) risk models[14] for patients undergoing AVR + (MVR or MVr)

TABLE 1 Number of Operations in the 3-Year Cohort

	All Operations (N = 1093 sites)		Included in Composite (N = 977 sites)	
Procedure	No.	% of Total	No.	% of Total
Overall	715333	100.0	668270	100.0
CABG	482285	67.4	448569	67.1
AVR	76632	10.7	72067	10.8
AVR + CABG	48149	6.7	45254	6.8
MV repair	37333	5.2	35708	5.3
MVR	31801	4.4	29953	4.5
MV repair + CABG	13065	1.8	12247	1.8
MVR + CABG	10828	1.5	10118	1.5
AVR + MVR	7241	1.0	6846	1.0
AVR + MV repair	3990	0.6	3743	0.6
AVR+MVR + CABG	2267	0.3	2138	0.3
AVR + MV repair + CABG	1742	0.2	1627	0.2

AVR, aortic valve replacement; CABG, coronary artery bypass grafting; MV, mitral valve; MVR, mitral valve replacement.

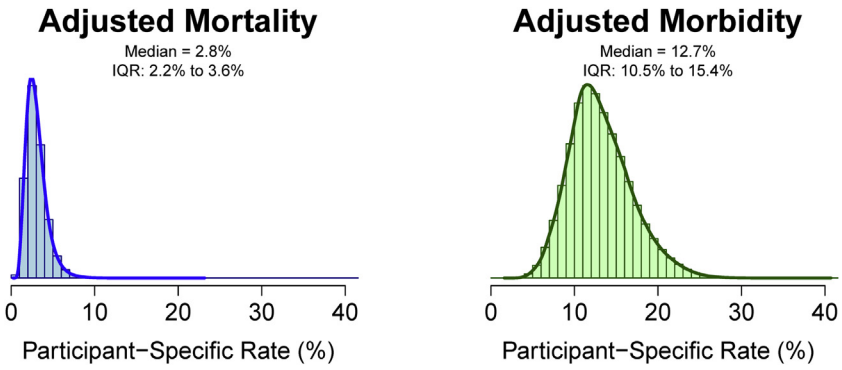

FIGURE 1 Estimated distributions of risk-adjusted mortality (blue) and morbidity (green) in the overall study populations. Median mortality was 2.8%, with an interquartile range (IQR) of 2.2% to 3.6%. Median morbidity was 12.7%, with an interquartile range of 10.5% to 15.4%. (Note: The height of each bar is proportional to the percentage of participants with composite scores in the interval depicted by the bar's position on the *X* axis.)

and AVR + (MVR or MVr) + CABG. To maximize calibration for the current study, all model coefficients were first reestimated using the current study's 3-year cohort.

PARAMETER ESTIMATION. Parameters of the bivariate random effects model were estimated in a Bayesian statistical framework by specifying a noninformative prior distribution for model parameters. The output of a Bayesian analysis is a probability distribution describing the relative likelihood of different estimates in light of the study data. An advantage of Bayesian estimation is the ability to express analysis results in terms of clinically relevant probabilities. For example, the model can provide an explicit quantification of the probability that a participant's risk-adjusted mortality or morbidity rate is higher or lower than the STS average. Posterior means and CrIs were calculated using Markov chain Monte Carlo (MCMC) simulations, as implemented in WinBUGS 3.4 software (MRC Biostatistics Unit, University of Cambridge).

COMPOSITE SCORES AND STAR RATINGS. The overall composite score was calculated for each participant as a weighted average of "1 minus the participant's risk-adjusted mortality rate" and "1 minus the participant's risk-adjusted morbidity rate." Rates were subtracted from 1 so that higher numbers imply better outcome performance. Mortality and morbidity rates were each weighted inversely by their respective SDs across participants.

After estimating composite scores for each participant, we used the 95% CrI around each participant's composite score to classify them as 1-star (worse than expected) participants if their 95% CrI was entirely lower than the STS average; 2-star (as expected) if their 95% CrIs overlapped the STS average; and 3-star (better than

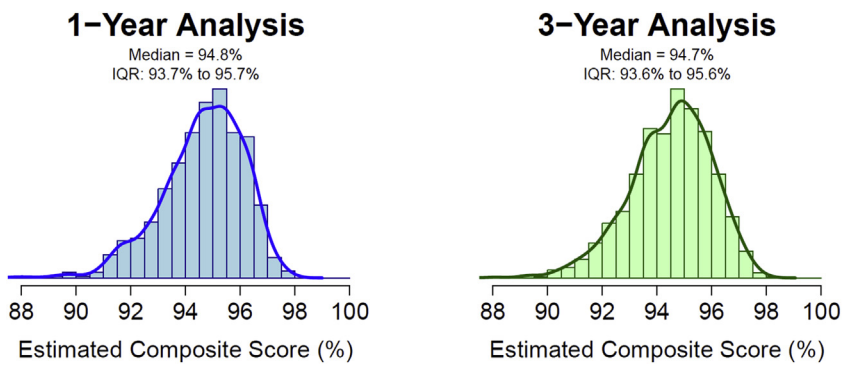

FIGURE 2 Distribution of composite scores in the study population on the basis of 1- and 3-year analytic cohorts. The overall distributions of estimated composite scores, and their medians and interquartile ranges (IQRs) were virtually identical for the 1-year (blue) and 3-year (green) study cohorts. (Note: The height of each bar is proportional to the percentage of participants with composite scores in the interval depicted by the bar's position on the *X* axis.)

expected) if their 95% CrI was entirely higher than the STS average.

Because of the design of the composite, each participant's overall score reflects their specific mix of patients and procedures.

COMPOSITE PROPERTIES. We performed several analyses to confirm that composite measure estimates behave as expected and have desirable statistical properties. To verify that each individual outcome exhibits adequate between-participant variation, we estimated the distributions of risk-adjusted mortality and morbidity across participants and displayed these results as a set of histograms. To assess face validity and confirm that the composite was not dominated by either single outcome, we compared the distribution of risk-adjusted outcome results for each individual outcome across star rating categories for the overall composite.

To ensure that the composite measure analysis had adequate statistical precision, we estimated the composite measure's signal-to-noise reliability, defined here as the proportion of variation in a measure that can be attributed to true signal variation as opposed to random statistical fluctuations. We performed several of these analyses separately for the 1-year and 3-year cohorts to determine the appropriateness of a 1-year vs a 3-year time window for future reporting of the composite measure.

Finally, a series of cross-tabulations was performed to compare the star ratings for the new multiprocedural composite with those obtained for individual procedures by using standard STS composite methodologies, with the caveat that the current STS CABG model uses only 1 year of data, whereas the multiprocedure composite and individual procedure composite measures use 3 years of data. Star ratings for the multiprocedure composite were compared with individual procedure composites in the subset of participants who had at least 100 eligible cases over 3 years in the multiprocedural composite analysis (n = 913). Each pairwise comparison was limited to participants who received a star rating in both analyses.

RESULTS

The composite development cohort for the 3-year analytic window included 668 270 procedures, with individual frequencies as shown in Table 1. Isolated CABG was by far the most common (n = 448 569; 67.1%), and isolated AVR (n = 72 067; 10.8%) had the second highest frequency. Overall, the procedures included in the composite measure encompassed 81% of a typical ACSD participant's practice. Corresponding procedure-mix data for the 1-year study cohort may be found in Supplemental Table 1.

TABLE 2 Star Rating Distributions and Reliabilities, 1- and 3-Year Samples

Sampling Period	Star Ratings 1-Star, n (%)	2-Star, n (%)	3-Star, n (%)	Reliability
1-y (N = 930)	74 (8)	697 (75)	159 (17)	0.69 (0.65-0.72)
3-y (N = 977)	157 (16)	570 (58)	250 (26)	0.81 (0.78-0.83)

Empirically derived weights for mortality and morbidity domains of the composite measures were 75% and 25%, respectively, in the 1-year analysis and 76% and 24%, respectively, in the 3-year analysis.

Figure 1 presents the estimated distributions of risk-adjusted mortality and morbidity for the overall study cohort, with medians of 2.8% and 12.7%, respectively. The distributions indicate substantial between-site variation not explained by case mix. Similarly, Figure 2 demonstrates the distributions of composite scores using 1-year and 3-year analytic cohorts; results were nearly identical regardless of the analytic window.

Table 2 shows the number and proportion of 1-, 2-, and 3-star participants and overall estimated measure reliability using 1-year and 3-year analysis cohorts. There was an 8-percentage point absolute increase in the proportion of 1-star programs when using 3-year vs 1-year data, a finding that corresponds to a doubling

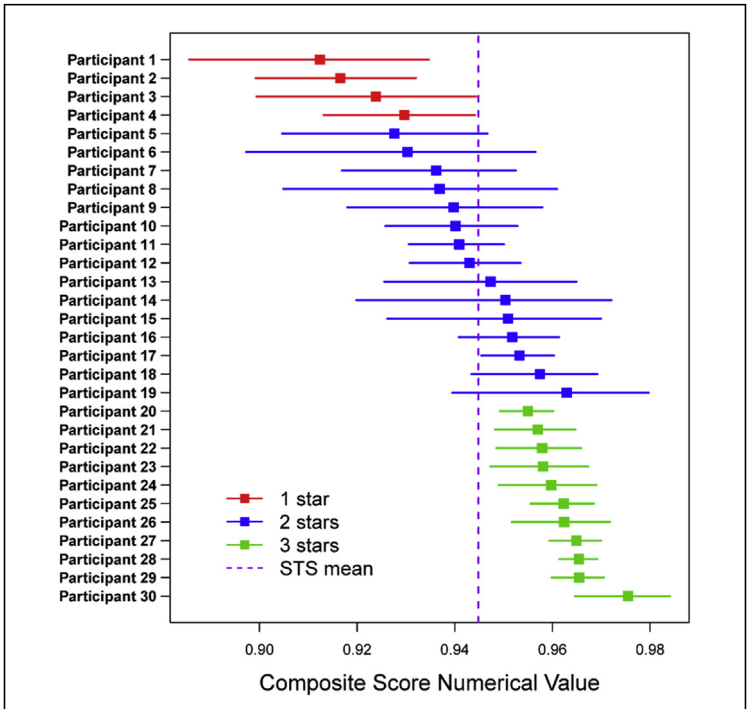

FIGURE 3 For illustrative purposes, Figure 3 shows the estimated multiprocedure composite CIs for 1-, 2-, and 3-star participants based upon a random sample of 30 participants. (STS, The Society of Thoracic Surgeons.)

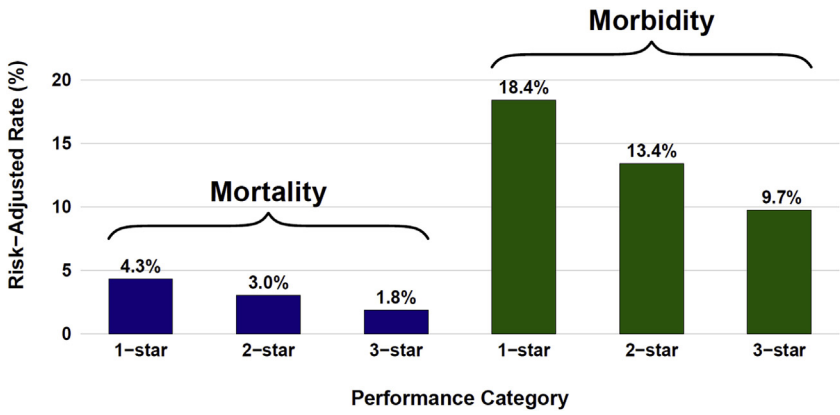

FIGURE 4 Adjusted mortality and morbidity rates for 1-, 2-, and 3-star participants. The risk-adjusted mortality (blue) and morbidity (green) rates for the Society of Thoracic Surgeons 1-, 2-, and 3-star participants demonstrated monotonic decreases as star ratings increased. For both risk-adjusted mortality and morbidity, average event rates for 3-star participants were roughly one-half those of 1-star participants.

(from 8% to 16%). Similarly, there was a 9-percentage point absolute increase in the proportion of 3-star participants using 3-year sampling, a 53% relative increase (from 17% to 26%). In addition to being able to classify a larger number of participants as 1 star or 3 star with a 3-year analysis window, average measure reliability also increased from 0.69 (0.65-0.72) to 0.81 (0.78-0.83).

For illustrative purposes, Figure 3 shows estimated CIs for 1-, 2-, and 3-star participants for the multiprocedure composite on the basis of a random sample of 30 participants.

Figure 4 shows the average unadjusted and adjusted mortality and morbidity rates for participants in the 1-star, 2-star, and 3-star rating categories, with monotonic decreases in adverse outcome rates moving from 1- to 3-star performance classifications. There were particularly notable differences when comparing 1-star and 3-star programs, with the latter having roughly one-half the rates of adverse outcomes as the former.

Additionally, a series of cross-tabulations (available on request) was performed to assess the correlations of mortality, morbidity, and overall composite scores. Among participants who had 3 stars for the overall composite, most (206 of 250) also had 3 stars for morbidity. The association between star ratings for mortality and the overall composite is less pronounced (126 of 250), perhaps reflecting the lower statistical power that is expected for mortality given the relative rarity of its occurrence. Participants who have 1 star for the overall composite have a mix of 1 or 2 stars for mortality and morbidity. Conversely, participants who have 3 stars for the overall composite have a mix of 2 or 3 stars for mortality and morbidity.

Table 3 demonstrates the distribution of 1-star, 2-star, and 3-star participants within different volume categories of eligible cases, by using 3 years of data and 95% CrIs. The lowest number (n = 4) and proportion (6%) of 1-star programs were in the lowest volume category (1-99 cases), which may appear contrary to what would have been expected on the basis of a volume-outcome association. However, this finding reflects the fact that at very low volumes, random sampling variation makes it difficult to categorize a provider as either a high or low performance outlier. Consequently, by far the most common performance classification (91%) among these very low volume providers is "as expected." There was no other consistent association of case volume and proportion of 1-star programs. A more consistent pattern, compatible with a volume-outcome association, was seen among 3-star, better-than-expected performing participants. Although only 2 of 64 participants with 1 to 99 cases (3%) and 9 of 143 participants (6%) in the 100- to 249-case group achieved a 3-star rating, there was a monotonic increase in the number and proportion of 3-star participants as the number of eligible cases increased, including nearly one-half of participants in the 750 or greater case volume group (151 of 307; 49%).

Table 4 shows estimated composite measure reliability above and below various case volume eligibility thresholds. STS generally requires average reliabilities of

TABLE 3 Performance Results by Volume Category (on the Basis of 95% Credible Intervals, 3-Year Analysis)

No. of Eligible Cases	No. of Participants	Worse Than Expected, n (%)	Same as Expected, n (%)	Better Than Expected, n (%)
All	977	157 (16)	570 (58)	250 (26)
1-99	64	4 (6)	58 (91)	2 (3)
100-249	143	27 (19)	107 (75)	9 (6)
250-499	280	59 (21)	182 (65)	39 (14)
500-749	183	32 (17)	102 (56)	49 (27)
750+	307	35 (11)	121 (39)	151 (49)

TABLE 4 Estimated Reliability at Various Reporting Thresholds (Eligible Cases)

Threshold	No. of Participants Meeting Threshold	Estimated Signal Variance	Reliability: Participant Volumes Higher Than Threshold	Reliability: Participant Volumes Lower Than Threshold
All Participants	977	0.0283	0.81 (0.78-0.83)	
≥50 eligible cases	950	0.0276	0.83 (0.81-0.85)	0.45 (0.16-0.71)
≥75 eligible cases	934	0.0272	0.84 (0.82-0.86)	0.45 (0.23-0.65)
≥100 eligible cases	913	0.0266	0.84 (0.82-0.86)	0.52 (0.35-0.67)
≥150 eligible cases	874	0.0261	0.85 (0.83-0.87)	0.59 (0.47-0.70)
≥200 eligible cases	823	0.0243	0.86 (0.84-0.87)	0.66 (0.57-0.73)
≥250 eligible cases	770	0.0231	0.86 (0.85-0.88)	0.68 (0.61-0.75)
≥500 eligible cases	490	0.0194	0.90 (0.88-0.91)	0.74 (0.70-0.78)
≥1000 eligible cases	196	0.0147	0.92 (0.90-0.94)	0.79 (0.75-0.81)

0.50 or greater for its performance measures. When necessary, STS will exclude participants with very low numbers of eligible cases from reporting to achieve at least 0.50 for average reliability. Results in Table 4 indicate that average reliability greater than 0.80 was achieved even when sites with relatively fewer eligible cases were included, although reliabilities were much lower for those few participants that failed to meet the lowest volume thresholds. For example, those with volumes lower than the 50-case threshold had average reliabilities of 0.45 (0.16-0.71), and those below the 75-eligible case threshold had average composite measure reliabilities of 0.45 (0.23-0.65). Only 64 of 977 total study participants (6.6%) failed to meet the 100 total procedures over 3 years volume threshold, below which average reliabilities would be less than 0.52; thus, roughly 93.4% of all STS participants would be eligible to receive an overall multiprocedure score if a threshold of 100 eligible cases was adopted.

Finally, cross-tabulations (Figure 5) of the star ratings for individual procedures vs the multiprocedural composite demonstrated that in each case, many programs are given 2 stars for the individual procedures but 1 or 3 stars for the multiprocedure composite, a rating that reflects a key strength of the latter. It has greater power to identify 1- and 3-star participants because it aggregates information across multiple different procedures within each participant to arrive at a relatively larger participant-specific sample size.

COMMENT

The STS QMTF has developed a multidimensional, multiprocedure, composite performance measure that is based on a 3-year analytic window. It includes risk-adjusted mortality and morbidity for 6 major categories of adult cardiac surgery procedures, accounting for approximately 81% of a typical STS participant's practice. Because of the large number of available end points, this comprehensive measure has high average reliability at all realistic levels of overall procedure volume. Using a threshold of 100 cases over 3 years, 94% of STS participants in the study population would be eligible to receive an STS composite score and rating.

This measure is designed to provide STS ACSD participants with a broad-based metric encompassing their overall performance. Because the measure is estimated from the observed and expected values for each patient who was cared for by the participant, it reflects the relative proportions of various types of procedures that the participant actually performs. Scores for participants focusing on coronary artery disease will be effectively weighted more by their greater proportion of CABG procedures, whereas scores for structural heart programs will have inherently more weight based upon their larger proportion of valve procedures.

This measure will be used by hospitals, cardiac surgery practice groups, and surgeons for quality improvement initiatives, internal hospital quality monitoring, regulatory compliance, and voluntary public reporting. Patients (eg, to assist in provider selection), payers (eg, for center of excellence designation), regulatory and government agencies (eg, oversight responsibilities), and others will appreciate the utility and simplicity of a single, comprehensive, methodologically sophisticated, and highly reliable measure that broadly reflects the overall performance of a participant in the STS ACSD.

As with other STS composite measures, a 1-star rating, particularly if consistent over multiple reporting periods, should alert the participant that a comprehensive review of their quality is needed, followed by appropriate improvement initiatives. Patients, payers, and regulators may also incorporate this information into their overall assessments of these participants regarding provider selection, reimbursement and preferred provider designation, and program certification, respectively.

This multidimensional composite measure is meant to augment and be used in parallel with the many procedure-specific composite measures in the STS quality measurement portfolio. Having both the multiprocedure and procedure-specific measures available will guide

	CABG composite		
Multiprocedure Composite	★	★★	★★★
★	19	126	0
★★	11	472	4
★★★	1	174	65

	AVR composite		
Multiprocedure Composite	★	★★	★★★
★	8	130	0
★★	5	448	4
★★★	0	188	48

	MV composite		
Multiprocedure Composite	★	★★	★★★
★	7	111	0
★★	7	391	1
★★★	1	180	46

	AVR + CABG composite		
Multiprocedure Composite	★	★★	★★★
★	11	123	0
★★	4	429	1
★★★	0	194	37

	MV + CABG composite		
Multiprocedure Composite	★	★★	★★★
★	4	29	0
★★	3	95	0
★★★	0	113	26

FIGURE 5 Star ratings for individual procedures vs multiprocedural composite. Cross-tabulations of the star ratings for individual procedures vs the multiprocedural composite were performed. These comparisons were limited to the subset of participants with at least 100 eligible cases over 3 years in the multiprocedural composite analysis (n = 913) and to participants receiving a star rating in both analyses. The number of participants with individual procedure star ratings (North America only) included 928 coronary artery bypass grafting operations (CABG), 880 aortic valve replacements (AVR), 782 mitral valve (MV) procedures, 846 AVR + CABG procedures, and 278 mitral + CABG procedures (relatively fewer participants because it has a minimum required sample size of ≥25 cases). The corresponding number of participants with both multiprocedural and individual procedure star ratings, and at least 100 eligible cases in the multiprocedural composite analysis (North America only), included 872 (of 913) for the multiprocedural vs CABG cross-tabulation, 831 (of 913) for the multiprocedural vs AVR cross-tabulation), 744 (of 913) for the multiprocedural vs mitral cross-tabulation, 799 (of 913) for the multiprocedural vs AVR + CABG cross-tabulation, and 270 (of 913) for the multiprocedural vs mitral + CABG cross-tabulation. Many 2-star programs for individual procedures are rated 1 or 3 stars for the multiprocedural composite. The latter has greater statistical power to identify low- and high-performing programs because it aggregates information across multiple different procedures, thus yielding larger sample sizes.

participant quality improvement activities and will provide patients with considerably greater information about the performance of prospective providers, both overall and for specific procedures of interest.

STUDY LIMITATIONS. A small percentage of STS participants will be ineligible to receive a multiprocedural composite score as a result of insufficient case volumes, most commonly because they are new programs. In that scenario, it would not be statistically or conceptually appropriate to assign an STS rating on the basis of a very small number of cases. It is extremely unlikely that any established program would fail to meet eligibility requirements.

Like any other measure based on indirectly standardized outcomes, the STS multiprocedure, participant-level composite measure should not be used to directly compare 1 participant with another because their overall mixes of cases may be quite different. This measure is correctly viewed as the performance of an STS participant for their specific mix of procedures and patients compared with what

would have been expected on the basis of results from the overall benchmark population of STS participants.

CONCLUSION. As the latest addition to its portfolio of multidimensional composite performance measures, the STS QMTF has developed a multiprocedural composite designed for use at the level of an STS ACSD participant. In conjunction with STS procedure-specific composite measures, this multiprocedural composite will provide all stakeholders with the most comprehensive information about both overall program performance and results for specific procedures.

This work was supported by internal STS funding.

REFERENCES

1. Shahian DM. Professional society leadership in health care quality: The Society of Thoracic Surgeons experience. *Jt Comm J Qual Patient Saf*. 2019;45:466-479.

2. Shahian DM, Grover FL, Prager RL, et al. The Society of Thoracic Surgeons voluntary public reporting initiative: the first 4 years. *Ann Surg*. 2015;262:526-535 [discussion: 33-35].

3. Shahian DM, Jacobs JP, Edwards FH, et al. The Society of Thoracic Surgeons National Database. *Heart*. 2013;99:1494-1501.

4. Shahian DM, Jacobs JP. Health services information: lessons learned from the Society of Thoracic Surgeons National Database. In: Sobolev B, Levy A, Goring S, eds. *Data and Measures in Health Services Research*. Springer US; 2016:1-24.

5. Shahian DM, Edwards FH, Ferraris VA, et al. Quality measurement in adult cardiac surgery: part 1–conceptual framework and measure selection. *Ann Thorac Surg*. 2007;83(suppl):S3-S12.

6. O'Brien SM, Shahian DM, DeLong ER, et al. Quality measurement in adult cardiac surgery: part 2–statistical considerations in composite measure scoring and provider rating. *Ann Thorac Surg*. 2007;83(suppl):S13-S26.

7. Shahian DM, He X, Jacobs JP, et al. The Society of Thoracic Surgeons isolated aortic valve replacement (AVR) composite score: a report of the STS Quality Measurement Task Force. *Ann Thorac Surg*. 2012;94:2166-2171.

8. Shahian DM, He X, Jacobs JP, et al. The STS AVR+CABG composite score: a report of the STS Quality Measurement Task Force. *Ann Thorac Surg*. 2014;97:1604-1609.

9. Badhwar V, Rankin JS, He X, et al. The Society of Thoracic Surgeons mitral repair/replacement composite score: a report of The Society of Thoracic Surgeons Quality Measurement Task Force. *Ann Thorac Surg*. 2016;101:2265-2271.

10. Rankin JS, Badhwar V, He X, et al. The Society of Thoracic Surgeons mitral valve repair/replacement plus coronary artery bypass grafting composite score: a report of The Society of Thoracic Surgeons Quality Measurement Task Force. *Ann Thorac Surg*. 2017;103:1475-1481.

11. Shahian DM, He X, Jacobs JP, et al. The Society of Thoracic Surgeons composite measure of individual surgeon performance for adult cardiac surgery: a report of The Society of Thoracic Surgeons Quality Measurement Task Force. *Ann Thorac Surg*. 2015;100:1315-1324 [discussion: 24-25].

12. 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;105:1411-1418.

13. O'Brien SM, Feng L, He X, et al. The Society of Thoracic Surgeons 2018 adult cardiac surgery risk models: part 2-statistical methods and results. *Ann Thorac Surg*. 2018;105:1419-1428.

14. Jacobs JP, Shahian DM, Badhwar V, et al. The Society of Thoracic Surgeons 2021 adult cardiac surgery risk models for multiple valve operations. *Ann Thorac Surg*. 2022;113:511-518.