The STS AV + CABG Composite Score: A Report of the STS Quality Measurement Task Force

David M. Shahian, MD, Xia He, MS, Jeffrey P. Jacobs, MD, J. Scott Rankin, MD, Karl L. Welke, MD, Fred H. Edwards, MD, Giovanni Filardo, PhD, MPH, Frank L. Fazzalari, MD, Anthony Furnary, MD, Paul A. Kurlansky, MD, J. Matthew Brennan, MD, MPH, Vinay Badhwar, MD, and Sean M. O’Brien, PhD

Department of Surgery and Center for Quality and Safety, Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts (DMS); Duke Clinical Research Institute, Durham, North Carolina (XH, JMB, SMO); All Children’s Hospital, John Hopkins University, St. Petersburg, Florida (JP); Centennial Medical Center, Vanderbilt University, Nashville, Tennessee (JSR); Children’s Hospital of Illinois and the University of Illinois College of Medicine, Peoria, Illinois (KFW); University of Florida College of Medicine, Jacksonville, Florida (FHE); Institute for Health Care Research and Improvement, Baylor Health Care System, Dallas, Texas (GF); University of Michigan Health System Cardiac Surgery Department, Crittenton Hospital, Rochester, Michigan (FLF); Starr-Wood Cardiac Group, Portland, Oregon (AF); Columbia HeartSource, New York, New York (PAK); and University of Pittsburgh Medical Center, Pittsburgh, Pennsylvania (VB)

Background. The Society of Thoracic Surgeons (STS) is developing a portfolio of composite performance measures for the most commonly performed adult cardiac procedures. This manuscript describes the third composite measure in this series, aortic valve replacement (AVR) combined with coronary artery bypass grafting surgery (CABG).

Methods. We identified all patients in the STS Adult Cardiac Surgery Database who underwent AVR + CABG during recent 3-year (July 1, 2009, through June 30, 2012) and 5-year (July 1, 2007, through June 30, 2012) periods. Variables from the STS risk model for AVR + CABG were used to adjust morbidity and mortality outcomes. Evidence for internal mammary artery use in AVR + CABG was examined. We compared composite measures constructed using 3 or 5 years of outcomes with Bayesian credible intervals of 90%, 95%, or 98%. The final STS AVR + CABG composite performance measure is based on 3 years of data and 95% credible intervals. It includes risk-adjusted mortality and morbidity but not internal mammary artery use.

Results. Median composite score is 91.0% (interquartile range, 89.5% to 92.2%). There were 2.6% (24 of 915) one-star (lower performing) and 6.5% (59 of 915) three-star (higher performing) programs. Morbidity and mortality decrease monotonically as star ratings increase. The percentage of three-star programs increased substantially among programs that performed more than 150 procedures over 3 years compared with those performing 25 to 50 procedures (32.8% versus 1.6%). Measure reliability was 0.51.

Conclusions. The STS has developed a composite performance measure for AVR + CABG based on 3-year data samples and 95% credible intervals. This composite measure identified 9.1% of STS participants as having higher or lower than expected performance.


The modern era of health care performance measurement began with the publication of inadequately adjusted mortality rates by the federal government in 1986 [1]. Because of methodological flaws, this particular initiative was widely criticized [2]. However, the concept of using outcomes such as mortality to assess health care quality has been progressively refined, including the use of risk adjustment and careful selection of appropriately homogeneous patient cohorts (eg, isolated aortic valve replacement [AVR]). These enhancements make comparisons among institutions more equitable, first by adjusting for inherent patient risk, and second, by assuring that programs performing higher risk combined procedures are not penalized by having these cases analyzed in the same category as isolated procedures.

A more recent evolution in health care performance measurement has been the recognition that mortality alone may not be an adequate measure of health care quality. For some conditions and procedures, deaths are infrequent and sample sizes are small, making it difficult to discriminate among providers. In other instances, mortality is an important end point, but there are other significant outcomes (eg, stroke, major infections, and dialysis-dependent renal failure) and process measures (eg, appropriate antibiotic administration) that could contribute to a more comprehensive understanding of the quality of care [3, 4].

In 2007, the Society of Thoracic Surgeons (STS) Quality Measurement Taskforce began developing a portfolio of procedure-specific composite performance measures that
embraced multiple dimensions of quality. First in this series was the STS composite measure for coronary artery bypass grafting (CABG), a four-domain composite consisting of risk-adjusted mortality; risk-adjusted occurrence of any of the five major complications of CABG (referred to as an “any or none” measure, in this case indicating the occurrence of stroke, reoperation, sternal infection, renal failure, or prolonged ventilation); use of at least one internal mammary artery (IMA) graft; and administration of all four National Quality Forum (NQF)–endorsed perioperative medications [3, 4]. In addition to giving providers feedback on their relative performance, this metric also has been used for voluntary public reporting [5], and it is now NQF-endorsed.

In 2012 the STS Quality Measurement Taskforce published and implemented a composite performance measure for isolated AVR [6]. Because NQF-endorsed process measures for AVR do not exist, and because there is an evolving strong preference for outcomes as opposed to process measures to evaluate health care quality, the STS AVR composite was intentionally designed to consist only of risk-adjusted mortality and morbidity. This report describes the development of the third in our series of composite performance measures, the STS AVR + CABG composite.

Material and Methods

Study Cohort

We first identified all patients in the STS Adult Cardiac Surgery Database (STS ACSD) who underwent AVR + CABG procedures during two periods: July 1, 2009, through June 30, 2012 (3 years) and July 1, 2007, through June 30, 2012 (5 years). All operations were subsequently included except those for which the STS risk scores could not be computed because of missing key variables (13 and 22 during 3- and 5-year study periods, respectively). For mortality analyses, in the rare instance that a patient was readmitted for another operation within 30 days of the previous surgery (13 patients [0.01%] in the 5-year sample, 4 patients [0.01%] in the 3-year sample), only the first operation was included to avoid potentially double counting mortality for the same patient. All admissions were included for the morbidity domain. As in our previous composite measures, STS participants (usually a hospital, occasionally a surgical group, and rarely an individual surgeon) with fewer than 10 eligible cases during the 3-year or 5-year study periods were not included in the corresponding analyses (137 [13.0%] participants excluded in the 3-year sample; 113 [10.4%] excluded in the 5-year sample).

Estimation of Composite Scores and Star Ratings

The statistical approaches used to estimate each participant’s composite AVR + CABG score and star rating were identical to those previously described for the STS CABG composite measure and the STS isolated AVR composite measure [3, 4, 6]. As in these previous measures, the scores for the two domains were assigned weights that were proportional to the reciprocal of their standard deviations. Variables from the published STS risk model [7] for AVR + CABG were used to adjust the morbidity and mortality outcomes.

For the AVR + CABG composite, it was possible to use some of, but not all, the existing performance domains from our previous composite measures for isolated AVR and isolated CABG. Regarding the AVR component, the previously mentioned constraints led to a decision to only use outcomes and not process measures. For the CABG component, inclusion of IMA use warranted careful consideration. After a thorough review of the evidence base (discussed subsequently), the Quality Measurement Taskforce determined that the evidence for IMA use in AVR + CABG procedures did not meet NQF measure endorsement criteria (eg, quantity, quality, and consistency of the evidence base) [8].

Time Frame and Credible Interval Determinations

Analyses used to evaluate 3- versus 5-year sample periods and 90%, 95%, or 98% Bayesian credible intervals (CrI) were also identical to our previously published methodologies [3, 4, 6]. Our goal was to have adequate sample sizes while at the same time recognizing that the inclusion of data from more remote time frames might make scores less relevant to current practice (eg, evolving surgical indications, technical innovations, new surgical staff). We also investigated various CrIs to classify participants as outliers. Here the objective was to identify the CrI with the highest specificity for determining true outliers, while also producing a reasonable degree of discrimination among providers.

Reliability Estimation

Finally, we used our previously described methods [6] to estimate reliability, the ratio of signal to noise, where signal is the proportion of variation in measured performance attributable to true differences and noise is random statistical variation [9]. We defined reliability as the squared correlation between the calculated AVR + CABG composite score and the “true” score, the latter estimated using Markov Chain Monte Carlo simulations.

Results

Table 1 shows the number of procedures, hospitals, and end points from our initial descriptive analyses, using 3-year and 5-year samples. The ultimate decision regarding time frame was made on practical grounds, as there are no absolute statistical criteria. Although 5-year samples resulted in more one- and three-star outliers, this was achieved at the cost of using much older data that might not reflect current practice at a given program (eg, different surgeons or techniques). As the 3-year samples provided patient cohorts and numbers of end points that were similar to those in our isolated AVR composite, and used more timely data than the 5-year samples, we chose the former.

Figure 1 plots the distribution of observed mortality and any-or-none morbidity rates across STS ACSD participants.
Table 2 shows the number of STS ACSD participants classified as high- (three-star), average- (two-star), or low- (one-star) performing outliers using 90%, 95%, or 98% Bayesian CrIs and 3-year or 5-year samples. Although there is no clear statistical criterion on which to base our choice of the most appropriate CrI (and corresponding Bayesian probability), from a practical perspective we believe that 98% CrIs and 3-year samples provided higher specificity but inadequate differentiation among programs. Conversely, although the percentage of high- and low-performing programs was substantially larger with 90% CrIs, this was thought to provide insufficient certainty about outlier status to assure face validity. Based on these practical rather than statistical considerations, we chose 95% CrIs (corresponding to 97.5% Bayesian probability), as in our previous isolated AVR composite measure. This yielded 2.6% (24 of 915) one-star and 6.5% (59 of 915) three-star programs, or nearly 10% overall high- or low-performing outliers. This seemed to be the best compromise, providing both acceptable probability of accurate outlier classification while still providing reasonable discrimination among programs.

As a sensitivity analysis, we estimated outliers using a 95% CrI and 3-year or 5-year data among the 915 participants from whom data were available from both time frames. In this subset, 89% of programs did not change ratings when moving to the longer sample time frame. This is comparable to the number of sites not changing classification when going to a longer time frame in the isolated AVR and also the isolated CABG composites.

The AVR + CABG composite performance measure score is a weighted average of two domain scores: risk-standardized mortality and risk-standardized morbidity. Table 3 shows the standard deviations and resulting weights of the two components of the AVR + CABG composite, using both 3-year and 5-year samples. Similar to the findings in our previous composites, mortality has approximately four times the weight of morbidity in determining the overall composite score because of its smaller standard deviation.

Figure 2 shows the distribution of composite scores using 95% CrIs and a 3-year sampling period. The median score was 91.0% and the interquartile range was 89.5% to 92.2%.

Table 4 shows the observed and risk-adjusted mortality and morbidity rates by star rating. For both sampling time frames and each of the three CrIs examined, morbidity and mortality rates declined progressively with increasing star ratings. This provides some evidence of the validity of these ratings—that is, the star ratings accurately reflect the clinical outcomes they are meant to represent. For example, for 3-year sampling period and using 95% CrIs, the average risk-adjusted mortality rates for one-, two-, and three-star programs were 7.8%, 4.5%, and 2.0%, respectively. Corresponding values for morbidity were 40.8%, 25.5%, and 15.8%, respectively.

Table 2. Star Ratings for Various Credible Intervals and Sample Periods

<table>
<thead>
<tr>
<th>Credible Interval</th>
<th>1-Star</th>
<th>2-Star</th>
<th>3-Star</th>
<th>1-Star</th>
<th>2-Star</th>
<th>3-Star</th>
</tr>
</thead>
<tbody>
<tr>
<td>90% CrI</td>
<td>43</td>
<td>780</td>
<td>92</td>
<td>87</td>
<td>742</td>
<td>139</td>
</tr>
<tr>
<td>95% CrI</td>
<td>24</td>
<td>832</td>
<td>59</td>
<td>83</td>
<td>818</td>
<td>97</td>
</tr>
<tr>
<td>98% CrI</td>
<td>8</td>
<td>871</td>
<td>36</td>
<td>32</td>
<td>863</td>
<td>73</td>
</tr>
</tbody>
</table>

CrI = credible interval.

Table 3. Composite Measure Component Weighting in 3- and 5-Year Samples

<table>
<thead>
<tr>
<th>Variable</th>
<th>3-Year Sample</th>
<th>5-Year Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard deviation</td>
<td>2.0</td>
<td>1.9</td>
</tr>
<tr>
<td>Weight</td>
<td>0.78</td>
<td>0.79</td>
</tr>
</tbody>
</table>

Table 1. Sample Sizes and Outcomes

<table>
<thead>
<tr>
<th>Variable</th>
<th>3-Year</th>
<th>5-Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>Participant (n)</td>
<td>915</td>
<td>968</td>
</tr>
<tr>
<td>Operation (n)</td>
<td>53,827</td>
<td>88,833</td>
</tr>
<tr>
<td>Mortality (n [%])</td>
<td>2,291 (4.3)</td>
<td>4,080 (4.6)</td>
</tr>
<tr>
<td>Any major morbidity (n [%])</td>
<td>13,152 (24.4)</td>
<td>22,381 (25.2)</td>
</tr>
<tr>
<td>Prolonged ventilation</td>
<td>9,765 (18.1)</td>
<td>16,548 (18.6)</td>
</tr>
<tr>
<td>Deep sternal infection</td>
<td>201 (0.4)</td>
<td>380 (0.4)</td>
</tr>
<tr>
<td>Permanent stroke</td>
<td>1,129 (2.3)</td>
<td>1,840 (2.3)</td>
</tr>
<tr>
<td>Renal failure</td>
<td>3,294 (6.4)</td>
<td>5,789 (6.9)</td>
</tr>
<tr>
<td>Reoperations</td>
<td>3,254 (6.0)</td>
<td>5,890 (6.6)</td>
</tr>
</tbody>
</table>

* Prolonged ventilation, deep sternal infection, permanent stroke, renal failure, and reoperations.

Fig 1. Distribution of risk-adjusted morbidity (top) and mortality (bottom) rates. (IQR = interquartile range.)
Table 5 shows the distribution of star ratings by volume category. At lower volumes, “average” scores predominate, as there is less information on which to confidently classify a provider as high- or low-performing outlier (although a few are identified). At progressively larger volumes, the percentage of high-performing, three-star programs increases substantially, particularly at volumes of more than 150 procedures during the 3-year period. This probably reflects both a positive volume-outcome association and increasing sample sizes on which to base inferences about relative performance. Conversely, the percentage of programs with one-star ratings is consistently less than 10% across the wide range of volumes. These findings are very similar to those we have previously reported for the STS isolated CABG composite score [10].

Reliability of the STS AVR + CABG measure using 3 years of data was 0.51 (95% CrI, 0.46 to 0.55). For comparison, reliability of the STS isolated CABG composite measures was 0.77 (95% CrI, 0.73 to 0.80) for the most recent 1 year of data. Similarly, reliability of the STS isolated AVR composite measure was 0.49 (95% CrI, 0.44 to 0.53) for the most recent 3 years of data.

Comment
We have described the development and operational characteristics of the third in a series of STS composite performance measures in cardiac surgery. The AVR + CABG composite includes both risk-adjusted mortality and risk-adjusted, any-or-none morbidity, but no process measures. The inclusion of IMA use, particularly for left anterior descending coronary artery (LAD) disease, was considered but ultimately rejected, as described in the following section.

Use of the Internal Mammary Artery in Aortic Valve Replacement–Coronary Artery Bypass Grafting Surgery
Use of the IMA is a component of our isolated CABG composite, and it is universally accepted as a “best practice” in isolated CABG surgery, particularly when the LAD requires grafting. To assess whether this measure was also appropriate for inclusion in the AVR + CABG composite, we examined the evidence supporting IMA grafting in the context of combined valve + CABG procedures, as opposed to isolated CABG from which most studies demonstrating IMA superiority have been derived.

Despite the abundance of empirical studies supporting use of the IMA for LAD grafting in isolated CABG [11-13], data specifically addressing IMA use in patients undergoing AVR + CABG are remarkably sparse. The most recent American College of Cardiology/American Heart Association guidelines for the management of patients with valvular heart disease assigns a class Ila recommendation (“is reasonable”) for IMA grafting of a
significantly diseased LAD, based on level of evidence C (ie, expert opinion) [14].

There has been a definite trend toward increased usage of the IMA for combined AVR + CABG procedures, from 0% utilization in early studies to 74% in more recent reports [15–21]. In the STS ACSD between 2010 and 2012, approximately 90% of AVR patients with left main or proximal LAD disease received an IMA graft. Most patients who did not receive an IMA had a documented reason (eg, unsuitable conduit).

Historically, studies have raised concerns regarding postoperative IMA flows compared with saphenous vein grafts (SVG) [22–25], and some have regarded this as a potential issue for IMA use in aortic stenosis patients with left ventricular hypertrophy. Other concerns have included potential increases in pulmonary morbidity or chest wall bleeding from IMA harvest, and wound healing problems. Conversely, favoring IMA use is the avoidance of SVG harvest and elimination of the need for a proximal anastomosis, especially in the “crowded” postaortotomy aorta. In those studies specifically reporting perioperative morbidity and mortality for IMA to LAD grafting during AVR (with varying degrees of risk adjustment), no statistically significant difference was observed between IMA and SVGs [17–20].

The data on long-term survival benefits of IMA grafting in valve + CABG are inconsistent. Although an initial report from their group showed no significant improvement in long-term survival with IMA versus SVGs [16], Herijgers and colleagues [26] performed a subset analysis of 535 patients younger than 70 years undergoing valve and CABG surgery. Within this subgroup, relative risk for late death was significantly reduced (risk ratio, 0.51; 95% confidence interval, 0.29 to 0.92; p = 0.025) with the use of the IMA. Tribouilloy and colleagues [27] from the Mayo Clinic found that for patients undergoing CABG in combination with replacement of left-side regurgitant valves, IMA grafting was an independent predictor of survival (hazard ratio, 0.57; p = 0.011). Gall and associates [18] reviewed outcomes of patients in the Duke registry who underwent AVR + CABG during the early years after the introduction of IMA grafting techniques (1984 to 1994), comparing patients with LAD disease who underwent IMA versus SVG. Use of IMA was associated with improved actuarial survival at 1 and 5 years (average follow-up, 41.2 months) and was an independent predictor of lower long-term mortality (relative risk, 0.57; p = 0.031) in a Cox proportional hazards model. In a study from the Cleveland Clinic, Beach and colleagues [28] identified use of the IMA as an independent predictor of improved long-term survival (regression coefficient - 0.24 ± 0.079; p = 0.002) in AVR + CABG patients, although the data were not stratified for LAD disease.

In contrast to these studies favoring IMA use, others have failed to demonstrate superiority. Karthik and colleagues [19] used propensity-score adjustment to compare survival in patients with comitant valvular surgery who also had either left IMA or SVG. After a mean follow-up of 3.7 years, left IMA grafting conferred no statistically significant reduction in the relative risk of mortality (relative risk, 0.91; 95% confidence interval, 0.61 to 1.37; p = 0.62). Among the subset of patients who underwent AVR + CABG, long-term survival was not significantly improved with left IMA grafting (adjusted relative risk, 1.05; 95% confidence interval, 0.69 to 1.60; p = 0.82). To avoid confounding issues related to structural failure of bioprosthetic valves, Jones and colleagues [20] examined patients undergoing AVR with mechanical valves and had 100% long-term follow-up. Compared with SVG, multivariable analysis did not find that use of IMA grafts significantly improved survival.

Consistency of findings in the evidence base is one major criterion for NQF endorsement of a performance measure [8]. Such consistency is not evident among studies of IMA use combined with AVR. Potential explanations include the limited number of patients in each study; different eras, patient age groups, and risk factors; and substantial variation in statistical techniques. In addition to these methodological differences, important clinical factors may obfuscate the impact of IMA use in AVR + CABG as opposed to isolated CABG procedures. For example, the ravages of longstanding valvular disease, including left ventricular hypertrophy, dilatation, or reduced ejection fraction, may dominate both short- and long-term outcomes, thus obscuring any potential benefit of IMA grafting. Conversely, in patients with extensive coronary artery disease, the importance of left IMA grafting to the LAD could emerge as the most significant predictor of long-term survival, even in the presence of coexisting valvular disease. Further complicating these studies is the inconsistent availability of information regarding valvular pathophysiology (stenosis versus regurgitation) as well as the location and severity of coronary lesions. Finally, the subtleties of mechanical versus biologic prostheses and the potential for either thromboembolic or anticoagulation-related events could also make it more difficult to evaluate superiority of the IMA.

In summary, although many experts believe that the IMA is the preferred conduit to graft a diseased LAD during combined AVR + CABG, there is currently an insufficient evidence base to include this as a performance measure. Furthermore, STS ACSD data suggest that this conduit is currently being used in more than 90% of AVR + CABG procedures in which the LAD is grafted, so a major performance gap does not exist.

Limitations
Participation in the STS ACSD is voluntary. However, 90% to 95% of all cardiac surgical programs nationally participate, and an independent external audit has demonstrated high data collection accuracy.

Process measures such as IMA use may warrant inclusion in subsequent versions of this measure if the evidence base evolves. There are also other outcomes measures that may be considered in the future (eg, patient-reported outcomes, failure to rescue) if there is sufficient evidence. The AVR + CABG composite measure is based entirely on short-term outcomes. Long-term outcomes are important to patients but are not currently captured in the STS database. Mortality and morbidity domains were
weighted equally in the composite (after rescaling each domain score to have the same standardization), but other weightings are possible. No single set of weights may reflect the subjective preferences of all possible users.

Conclusions
The STS has developed a composite performance measure for AVR + CABG based on two domains, risk-adjusted mortality and the risk-adjusted occurrence of any of five major complications. The STS will continue to develop and implement additional composite performance measures, not only in adult cardiac surgery but also in congenital and general thoracic surgery. Our ultimate goal is to construct a portfolio of measures that encompasses all the major procedures performed in our specialty. This will become the most comprehensive and scientifically accurate source of performance metrics for cardiothoracic surgery and will be available for use by all interested stakeholders. This initiative may serve as a model for other medical and surgical specialties interested in performance measurement.

References