Association of hospital coronary artery bypass volume with processes of care, mortality, morbidity, and the Society of Thoracic Surgeons composite quality score

David M. Shahian, MD, a Sean M. O’Brien, PhD, b Sharon-Lise T. Normand, PhD, c Eric D. Peterson, MD, MPH, b and Fred H. Edwards, MD d

Objective: This study examines the association of hospital coronary artery bypass procedural volume with mortality, morbidity, evidence-based care processes, and Society of Thoracic Surgeons composite score.

Methods: The study population consisted of 144,526 patients from 733 hospitals that submitted data to the Society of Thoracic Surgeons Adult Cardiac Database in 2007. End points included use of National Quality Forum–endorsed process measures (internal thoracic artery graft; preoperative β-blockade; and discharge β-blockade, antiplatelet agents, and lipid drugs), operative mortality (in-hospital or 30-day), major morbidity (stroke, renal failure, reoperation, sternal infection, and prolonged ventilation), and Society of Thoracic Surgeons composite score. Procedural volume was analyzed as a continuous variable and by volume strata (<100, 100–149, 150–199, 200–299, 300–449, and ≥450). Analyses were performed with logistic and multivariate hierarchical regression modeling.

Results: Unadjusted mortality decreased across volume categories from 2.6% (<100 cases) to 1.7% (>450 cases, P < .0001), and these differences persisted after risk factor adjustment (odds ratio for lowest- vs highest-volume group, 1.49). Care processes and morbidity end points were not associated with hospital procedural volume except for a trend (P = .0237) toward greater internal thoracic artery use in high-volume hospitals. The average composite score for the lowest volume (<100 cases) group was significantly lower than that of the 2 highest-volume groups, but only 1% of composite score variation was explained by volume.

Conclusion: A volume–performance association exists for coronary artery bypass grafting but is weaker than that of other major complex procedures. There is considerable outcomes variability not explained by hospital volume, and low volume does not preclude excellent performance. Except for internal thoracic artery use, care processes and morbidity rates were not associated with volume. (J Thorac Cardiovasc Surg 2010;139:273-282)

Subsequent to the pioneering work of Luft et al,1 volume has commonly been used as a structural measure of health care quality, particularly when more direct measures are either unavailable or unreliable because of small sample sizes. Although a volume–mortality association of varying strength has been observed for many procedures, including coronary artery bypass grafting (CABG) procedures, few studies have examined the associations between volume and more comprehensive measures of provider performance. Furthermore, even more recent studies of the CABG volume–mortality association are based on data that are now nearly a decade old2,3 and might not reflect current practice, particularly given the marked decrease in average CABG volume at most hospitals.

This study uses contemporary, national clinical data to examine the association between hospital CABG volume and the Society of Thoracic Surgeons (STS) CABG composite score, as well as its 4 performance domains: (1) operative mortality; (2) major morbidity; (3) operative care; and (4) perioperative medications. The STS CABG composite score has been in use since 2007 and is arguably the most comprehensive, multidimensional measure of CABG performance currently available.4,5

MATERIALS AND METHODS

Study Population
The study population consists of adults (age, 20–100 years) who underwent isolated CABG surgery at an STS Adult Cardiac Database (ACDB)–participating hospital between January 1, 2007, and December 31, 2007. Among 154,210 eligible CABG operations, we excluded 415

From the Department of Surgery and Center for Quality and Safety, a Massachusetts General Hospital, Harvard Medical School, Boston, Mass; the Duke Clinical Research Institute, b Durham, NC; the Department of Health Care Policy, Harvard Medical School, and the Department of Biostatistics, c Harvard School of Public Health, Boston, Mass; and the Division of Cardiothoracic Surgery, d University of Florida, Gainesville, Fla.

Disclosures: None.


Received for publication May 6, 2009; revisions received Aug 14, 2009; accepted for publication Sept 2, 2009; available ahead of print Dec 21, 2009.

Address for reprints: David M. Shahian, MD, Massachusetts General Hospital, Department of Surgery and Center for Quality and Safety, 55 Fruit St, Boston, MA 02114 (E-mail: dshahian@partners.org).

0022-5223/36.00
Copyright © 2010 by The American Association for Thoracic Surgery
doi:10.1016/j.jtcvs.2009.09.007
cases for which the hospital could not be determined and an additional 8854 records from 103 hospitals that submitted data for only part of 2007. Three hospitals were excluded (n = 384 patients) because of high frequency (>5%) of missing data for National Quality Form (NQF) medication measures. Finally, we excluded 7 patients with missing sex and 24 patients whose data were submitted in a format other than STS version 2.52. The final study population consists of 144,526 patients from 733 hospitals.

There were several analysis-specific exclusion criteria: (1) patients with preoperative renal failure were excluded when analyzing postoperative renal failure; (2) patients with preoperative cerebrovascular accidents were excluded when analyzing stroke; (3) patients with prior CABG were excluded when analyzing internal thoracic artery (ITA) use; and (4) patients who died in the hospital were excluded when analyzing discharge medications.

**Unit of Inference**

The unit of inference for this study of CABG volumes is a hospital. This is a departure from some previous STS analyses based on “STS participant.” Hospital was chosen over participant because the latter can consist of multiple hospitals and surgeon groups. Hospital volume was also chosen to be consistent with existing volume–outcome studies and because it might be more relevant to the study of system-level effects.

**Hospital Identification**

Hospitals were identified by using the participant ID, hospital name, hospital ZIP code, and hospital state database fields. Ambiguous hospital names were adjudicated by using information from hospital Web sites and direct communication with STS data managers at participating hospitals.

**Primary Independent Variable: CABG Volume**

Hospital CABG volumes were calculated based on the 2007 calendar year after preliminary studies showed near-perfect ($r = 0.994$) correlation between 1-year (2007) and 2-year (2006–2007) volumes. For the purpose of defining hospital volume, all cases with CABG were counted, including CABG performed in combination with other cardiac or noncardiac procedures. The presumption was that all CABG cases, even when combined with other procedures, still afforded experience in CABG techniques. However, after determining total CABG case volumes, all subsequent analyses were restricted to patients undergoing isolated CABG. Patients undergoing concomitant procedures (eg, valve surgery) were excluded.

Analyses were performed with volume both as a continuous and categorical variable. Cut points for the latter (<100, 100–149, 150–199, 200–299, 300–449, and ≥450 total CABG cases) were selected after considering both previous studies and the empiric distribution of hospital volumes in our study population.

**Primary End Points**

Study end points include the overall STS composite CABG score, as well as the 4 performance domains and 11 individual NQF-endorsed measures from which it is derived: (1) all-cause operative mortality (in-hospital death regardless of timing and 30-day death regardless of venue); (2) major morbidity (both individual outcomes and any-or-none occurrence of the followings: renal failure, re-exploration, stroke, sternal infection, or prolonged ventilation); (3) failure to use at least 1 ITA bypass graft conduit; and (4) failure to use recommended perioperative medications (preoperative β-blockade and discharge antiplatelets, lipid-lowering agents, and β-blockade) both individually and as part of an all-or-none medication domain. Definitions of these end points were consistent with the report of the STS Quality Measurement Task Force report on composite measures$^4,5$ and with version 2.52 of the STS ACDB.

**Model Covariates**

Covariates for the patient-level hierarchical logistic regression models included all variables in the STS 2008 CABG mortality model$^6$ plus month of surgical intervention (to account for secular declining trends in mortality, which were small during the 1-year study period).

**Analyses**

The number of hospitals and patients, the distribution of preoperative patient characteristics, and unadjusted performance rates were determined for each 2007 volume category.

Random-effects logistic regression was performed with SAS GLIMMIX (SAS Institute, Inc, Cary, NC) to compare the frequency of adverse outcomes across volume categories, accounting for case mix and clustering. Odds ratios and confidence intervals (CIs) were calculated for each adverse outcome and volume category relative to the performance of hospitals with 450 or more CABG procedures. Unlike the method originally used to develop the STS composite score,$^4$ these models were fit by using individual patient-level data and not aggregated to the level of hospitals.

Random-effects logistic regression models were also fit using volume as a continuous rather than categorical variable. To allow for possible nonlinear effects, volume was modeled by using restricted cubic splines with a reference point of 500 CABG procedures and knots at 150, 300, and 450. Sensitivity to the choice of knot locations was assessed by refitting the models with knots at various percentiles of the empirical distribution of hospital volume. The shape of the estimated volume–outcome association was generally consistent across these different model specifications, and therefore only the first set of results are presented. Seven hospitals with volumes in excess of 1000 CABG procedures are not included in these plots, but their data were included in the analyses.

A multivariate hierarchical random-effects model for the composite CABG score was estimated by using WinBUGS software (Appendix 1).$^7$ The reported estimates are based on the Bayesian posterior distribution and are accompanied by Bayesian 95% credible intervals.

Using previously described methodology,$^4,5$ the STS star rating was determined for each hospital, and the scores were plotted against the hospital volume category. A 1-star rating indicates a 99% Bayesian probability that a hospital’s composite score is lower than the STS average (2-star rating), and a 3-star rating indicates a 99% Bayesian probability that a hospital’s composite score exceeds the STS average.

**Missing Data**

Missing data were managed by using multiple imputation, as implemented in SAS PROC MI with the MCMC option. We assumed data were missing at random, implying that conditional on all the observed data available to us, the missing data mechanism did not depend on the actual value of the missing variable.

**RESULTS**

Table 1 demonstrates the distribution of hospitals and isolated CABG procedures among 6 hospital volume...
TABLE 1. Distribution of hospitals, patients, and unadjusted performance results by volume category

<table>
<thead>
<tr>
<th>Hospital volume category</th>
<th>All</th>
<th>&lt;100</th>
<th>100–149</th>
<th>150–199</th>
<th>200–299</th>
<th>300–449</th>
<th>≥450</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hospitals</td>
<td>733 (100)</td>
<td>134 (18.3)</td>
<td>128 (17.5)</td>
<td>108 (14.7)</td>
<td>157 (21.4)</td>
<td>114 (15.6)</td>
<td>92 (12.6)</td>
</tr>
<tr>
<td>Patients</td>
<td>144,526 (100)</td>
<td>8056 (5.6)</td>
<td>12,740 (8.8)</td>
<td>14,789(10.2)</td>
<td>30,209(20.9)</td>
<td>31,585(21.9)</td>
<td>47,147(32.6)</td>
</tr>
</tbody>
</table>

Domain results

<table>
<thead>
<tr>
<th></th>
<th>% (95% CI)</th>
<th>% (95% CI)</th>
<th>% (95% CI)</th>
<th>% (95% CI)</th>
<th>% (95% CI)</th>
<th>% (95% CI)</th>
<th>P value for trend</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mortality</td>
<td>2.0 (1.9–2.1)</td>
<td>2.3 (2.0–2.6)</td>
<td>2.1 (1.8–2.4)</td>
<td>2.2 (2.0–2.4)</td>
<td>2.0 (1.9–2.2)</td>
<td>1.7 (1.5–1.9)</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Morbidity</td>
<td>13.8 (13.3–14.3)</td>
<td>14.8 (13.0–15.4)</td>
<td>13.9 (13.0–14.9)</td>
<td>13.9 (13.0–14.8)</td>
<td>14.0 (13.0–15.0)</td>
<td>13.2 (12.2–14.3)</td>
<td>.1129</td>
</tr>
<tr>
<td>ITA failure</td>
<td>6.1 (5.7–6.5)</td>
<td>6.8 (5.7–8.0)</td>
<td>6.5 (5.3–7.7)</td>
<td>6.5 (5.7–7.3)</td>
<td>5.4 (4.6–6.2)</td>
<td>5.8 (4.9–6.7)</td>
<td>.0456</td>
</tr>
<tr>
<td>Med failure</td>
<td>36.7 (35.5–38.0)</td>
<td>38.4 (33.0–38.0)</td>
<td>35.5 (34.0–39.4)</td>
<td>36.7 (34.1–38.8)</td>
<td>36.5 (33.0–38.5)</td>
<td>37.5 (34.8–40.3)</td>
<td>.7672</td>
</tr>
</tbody>
</table>

Morbidity measures

<table>
<thead>
<tr>
<th></th>
<th>% (95% CI)</th>
<th>% (95% CI)</th>
<th>% (95% CI)</th>
<th>% (95% CI)</th>
<th>% (95% CI)</th>
<th>% (95% CI)</th>
<th>P value for trend</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prolonged ventilation</td>
<td>9.2 (8.7–9.7)</td>
<td>9.9 (8.8–11.0)</td>
<td>9.7 (8.6–10.8)</td>
<td>9.1 (8.1–10.0)</td>
<td>9.3 (8.5–10.1)</td>
<td>9.4 (8.4–10.4)</td>
<td>8.8 (7.8–9.9)</td>
</tr>
<tr>
<td>DSWI</td>
<td>0.3 (0.3–0.4)</td>
<td>0.3 (0.2–0.4)</td>
<td>0.5 (0.3–0.6)</td>
<td>0.3 (0.2–0.4)</td>
<td>0.3 (0.2–0.3)</td>
<td>0.4 (0.3–0.5)</td>
<td>0.4 (0.3–0.4)</td>
</tr>
<tr>
<td>Stroke</td>
<td>1.0 (0.9–1.0)</td>
<td>0.9 (0.7–1.2)</td>
<td>0.9 (0.7–1.1)</td>
<td>1.1 (0.9–1.3)</td>
<td>1.0 (0.8–1.1)</td>
<td>1.0 (0.9–1.1)</td>
<td>0.9 (0.8–1.1)</td>
</tr>
<tr>
<td>Renal failure</td>
<td>3.0 (2.9–3.2)</td>
<td>3.4 (2.9–3.9)</td>
<td>3.1 (2.6–3.5)</td>
<td>3.1 (2.7–3.5)</td>
<td>3.1 (2.8–3.5)</td>
<td>3.2 (2.9–3.6)</td>
<td>2.7 (2.4–3.0)</td>
</tr>
<tr>
<td>Reoperation</td>
<td>5.1 (4.9–5.3)</td>
<td>5.4 (4.7–6.1)</td>
<td>5.1 (4.4–5.7)</td>
<td>5.1 (4.6–5.6)</td>
<td>5.0 (4.5–5.4)</td>
<td>5.2 (4.8–5.7)</td>
<td>5.0 (4.5–5.5)</td>
</tr>
</tbody>
</table>

Medication failure

<table>
<thead>
<tr>
<th></th>
<th>% (95% CI)</th>
<th>% (95% CI)</th>
<th>% (95% CI)</th>
<th>% (95% CI)</th>
<th>% (95% CI)</th>
<th>% (95% CI)</th>
<th>P value for trend</th>
</tr>
</thead>
<tbody>
<tr>
<td>Preop β-blockade</td>
<td>22.2 (21.3–23.2)</td>
<td>22.3 (20.3–24.3)</td>
<td>21.8 (20.1–23.5)</td>
<td>22.0 (20.2–23.8)</td>
<td>22.6 (20.9–24.3)</td>
<td>21.1 (18.9–23.3)</td>
<td>23.0 (20.9–25.1)</td>
</tr>
<tr>
<td>D/C antiplatelet</td>
<td>4.3 (3.7–5.0)</td>
<td>5.7 (3.3–8.0)</td>
<td>3.7 (3.0–4.3)</td>
<td>3.3 (2.8–3.9)</td>
<td>4.0 (3.4–4.5)</td>
<td>4.2 (3.6–4.8)</td>
<td>4.9 (3.0–6.7)</td>
</tr>
<tr>
<td>D/C β-blockade</td>
<td>9.7 (9.0–10.4)</td>
<td>11.3 (8.8–13.8)</td>
<td>9.7 (8.3–11.1)</td>
<td>9.7 (8.4–10.9)</td>
<td>9.6 (8.5–10.7)</td>
<td>9.1 (8.0–10.2)</td>
<td>9.9 (8.0–11.8)</td>
</tr>
<tr>
<td>D/C antilipids</td>
<td>11.6 (10.6–12.5)</td>
<td>13.6 (10.9–16.3)</td>
<td>10.8 (9.2–12.5)</td>
<td>11.8 (9.9–13.6)</td>
<td>11.2 (9.6–12.9)</td>
<td>11.5 (10.0–13.1)</td>
<td>11.5 (9.3–13.7)</td>
</tr>
</tbody>
</table>

CI, Confidence interval; ITA, internal thoracic artery; DSWI, deep sternal wound infection; D/C, discharge.

categories. Of the 733 STS-participant hospitals in this 2007 study cohort, 87.4% of hospitals performed fewer procedures than the 450 CABG case threshold recommended by Leapfrog, and 67.4% of the procedures were performed by hospitals with volumes of less than this threshold.

A complete table of patient characteristics for each volume category is available on request. For most characteristics, the absolute differences among volume categories were small, although most were statistically significant because of the large sample size. Several were of both statistical and clinical interest. For example, the frequency of Hispanic and Asian patients was significantly greater at the lowest-volume programs compared with that at the highest-volume programs (11.2% vs 3.7%, P < .001). Reoperative CABG was more prevalent at the largest compared with the smallest programs (4.5% vs 2.8%, P < .001), whereas patients in cardiogenic shock were more common at the smallest programs (2.5% vs 1.8%, P < .001). The latter might reflect the difficulty in transferring such critically ill patients to major tertiary centers in a timely fashion. The use of preoperative intra-aortic balloon pumps was higher in the smallest compared with the largest programs (9.9% vs 6.6%, P < .001), and this could reflect either practice style or their slightly increased prevalence of cardiogenic shock and emergency patients.

Table 1 also presents the unadjusted frequency of adverse outcomes and process noncompliance stratified by hospital volume category. The 2007 mortality rate (2.6%) for the lowest-volume category (<100 cases) was only slightly higher than the mortality (2.4%) for the highest-volume group (≥450 cases) in the most recent previous STS study. These findings are consistent with recently reported data from the Nationwide Inpatient Sample showing progressive improvements in CABG outcomes over time.

Mortality varied inversely with hospital volume group, with an absolute mortality difference of 0.9% between the lowest-volume (2.6%) and highest-volume (1.7%) categories (P for trend < .0001). The absolute difference in failure to use an ITA was 1% between the highest- and lowest-volume categories (5.8% vs 6.8%, P trend = .0456). There was no statistically significant difference in overall morbidity or perioperative medication use. The frequency of
postoperative renal failure decreased significantly with increasing hospital volume ($P = .0314$).

Figure 1 includes scatterplots of the unadjusted performance domain results for each of the 733 hospitals, using volume as a continuous rather than categorical variable. Much of the outcomes variability at lower-volume programs in all 4 plots is due to the sampling variation, a phenomenon that is the basis for so-called “funnel plots.”

Table 2 shows the results of random-effects logistic regression modeling accounting for clustering and case mix. Compared with the results for the 450 or greater case group, the adjusted odds ratios for mortality were inversely related to volume and were statistically significant for each volume category except 150 to 199 procedures. The largest adjusted odds ratio was in the lowest-volume category (1.49; 95% CI, 1.24–1.80). For nonfatal end points and process measure failures, only the odds ratio for renal failure in hospitals performing less than 100 CABG procedures was significantly greater than 1. There was a trend toward more consistent ITA use in higher-volume hospitals ($P = .0237$), but the odds ratios for all individual volume categories included 1. In a separate analysis of the 134 lowest-volume hospitals, there was no difference between hospitals with above- and below-average mortality in failure to use an ITA (6.7% vs 6.9%, $P = .55$) or failure to use appropriate medications (37.7% vs 39.1%, $P = .69$).

Figure 2 shows the results of random-effects logistic regression modeling with volume as a continuous variable instead of a categorical variable. The plots of medication compliance and overall morbidity are relatively horizontal, and the CIs encompass an odds ratio of 1. ITA failure has an inflection point at approximately 200 cases, but the CI only exceeds 1 at extremely low volume. The odds ratio for mortality increases monotonically with decreasing volume, with an increase in slope at approximately 200 procedures.

Figure 3 presents the mean STS composite scores with Bayesian 95% credible intervals for each volume category. The entire CI of the lowest-volume category lies substantially below the lower limits of the CIs for the 2 highest-volume categories. However, within each volume stratum, there is considerable variability in composite scores, and the overall variability...
proportion of variation in composite score explained by volume was only 1%.

Figure 4 is a scatterplot of the composite scores of all 733 hospitals plotted against volume as a continuous rather than categorical variable. Although composite scores are rather symmetrically distributed around the mean for programs of all sizes, the lowest scores occurred mainly among the lower-volume programs.

Figure 5 depicts the star ratings calculated by STS for all CABG composite score recipients. Smaller programs are less likely to achieve either 1- or 3-star status, primarily because there is less certainty about their true underlying performance. As volume increases, the number of 3-star programs increases, both because their results become more reliable and because their performance is higher on average. The number of 1-star programs remains relatively static.

**DISCUSSION**

**Volume and Mortality**

Numerous reports over the past 3 decades have documented the existence of an inverse association between procedural volume (surgeon, hospital, or both) and mortality, at least for more complex procedures. The magnitude and shape of these associations vary by procedure, as do the threshold levels for optimal performance.

CABG is performed more frequently than any other major complex operation, and it has been the subject of numerous volume–mortality studies. Compared with other complex but less frequently performed procedures, such as esophagectomy and pancreatocystectomy, most CABG studies have shown a weaker or absent volume–outcome association. Studies from New York have consistently shown a significant inverse association, which might be related to its long history of public reporting and limited number of very low-volume programs. Conversely, in California, a state with many low-volume hospitals and higher overall CABG mortality, a CABG volume–outcome association was no longer apparent after public report cards were introduced, suggesting that all hospitals were stimulated to improve performance.

Although an overall volume–mortality association does exist for CABG, there remains substantial mortality variability within every volume stratum, especially at very low volumes. High volume does not guarantee better outcomes at any specific program, nor does low volume preclude excellence.

As CABG volumes have decreased because of better medical therapy and the increasing use of percutaneous coronary intervention, there has been concern that outcomes would deteriorate because of diminishing experience. The magnitude and shape of these associations vary by procedure, as do the threshold levels for optimal performance.

**Volume and Use of Optimal Practices**

Empirical studies of the association between volume and use of evidence-based care processes are less numerous than those for volume and mortality. The clinical importance of an association between volume and process compliance is also confounded by inconclusive evidence relating the latter to outcomes.
A few surgical studies suggest that superior processes of care might be used by higher-volume surgeons and hospitals, although it is often difficult to differentiate the relative contributions of volume versus specialty training and practice.\textsuperscript{30–38}

Several clinical studies affirm the importance of optimal operative and perioperative care processes in providing high-quality CABG care.\textsuperscript{39–43} Discharge prescription of antiplatelet agents, lipid-lowering drugs, and $\beta$-blockade is based on innumerable studies demonstrating their effectiveness in secondary prevention of atherosclerotic progression.

Implications of the Current Study

The CABG volume–mortality association in this large, nationally representative sample remains relatively weak compared with that of other complex procedures, despite progressively lower average CABG volumes. As a group, the lowest-volume hospitals had significantly higher mortality rates. However, there was wide residual variability in outcomes at all hospital volumes. Because of this variation, volume alone is a poor predictor of an individual hospital’s true risk-adjusted mortality rate.

This study adds several new dimensions to our understanding of various CABG volume associations. First, there was no significant association between use of measured, evidence-based processes of care and program volume. The only exception was the significant $P$ value (for trend) of the adjusted odds ratios for ITA nonuse, although the 95\% CIs of the odds ratios for all individual volume categories included 1. This suggests that these measured CABG processes are applicable across the broad spectrum of program volumes, including small programs. In fact, it might be the widespread application of such care processes that explains the excellent results achieved by many lower-volume CABG programs in the United States, Japan, and elsewhere.\textsuperscript{10,12,13,44}

Given that a CABG volume–mortality association was observed in this study, one might have expected that this would be accompanied by a parallel association between process compliance and volume. Although this was not observed, it does not diminish the importance of using optimal
processes. The 4 NQF-endorsed measures we examined in this study represent only a subset of the universe of CABG care processes, which is generally true of most guidelines.45,46 Our data suggest that other unmeasured processes of care, including overall team functioning, might be more responsible for the volume–mortality association than the 4 NQF-endorsed measures we examined.

Our study is also one of the first to examine the association of procedural volume and the incidence of major complications. The overall unadjusted occurrence of any major complication varied little among all but the smallest and largest CABG volume categories, and even for this most extreme comparison, the spread was only 1.6% (13.2% vs 14.8%). With the exception of renal failure in the smallest programs, adjusted multivariable odds ratios revealed no statistically significant association of any complications with volume. The fact that complication rates did not vary by CABG volume, whereas mortality did demonstrate a volume association, suggests that larger programs might “rescue” patients more effectively from the sequelae of major complications.47

This is the first study of the association between hospital CABG volume and performance on the multidimensional STS composite performance score.4,5 Many stakeholders regard this as the most sensitive and comprehensive overall measure of provider CABG quality. STS studies have shown that it is much more discriminating of quality differences among providers than mortality alone, although mortality is a major driver of composite scores because of standardization used in composite score construction.4 Average CABG composite scores in this study were significantly lower for the lowest-volume group compared with the 2 highest-volume groups. Analyzing volume as a continuous variable, the lowest hospital composite scores were associated with very low-volume programs, and no higher-volume hospitals had such low scores.

Finally, our study demonstrates that there are fewer low-volume programs that achieve either better or worse than average star ratings using the STS composite, primarily because there is insufficient sample size to reliably differentiate their performance. Interestingly, the percentage of high-performing programs increases steadily above the 200 to 299 CABG case category, whereas the percentage of low-performing programs remains relatively constant. Within each volume stratum, there are about 10% to 15% of programs with composite performance below that of the STS average.

Limitations
Although broadly representative of national practice and encompassing a large majority of US programs, the STS ACDB is voluntary. It is possible that some low-volume, underperforming programs chose not to participate, which could skew the results of our analyses. Nonetheless, the high penetration (roughly 80% in 2007 and currently >90%) of the STS database among US cardiac surgery programs argues against this as a major concern.

Despite our best efforts, it is possible that some STS hospital CABG volumes could be inaccurate because this study required manual reclassification of STS participant data to the hospital level. A participant is often a hospital, but occasionally more than 1 STS participant surgeon or group might
practice at one institution. Overall, any volume inaccuracies in this study, which is based on a robust clinical data registry encompassing all age groups and payers, are likely to be less than those for studies derived from administrative data.

Only a small proportion of STS participant programs are audited annually, but the results of such audits have generally shown good overall coding accuracy. For most hospital CABG volume categories, about 90% of 30-day deaths in the STS database occur in the hospital, and these deaths are captured with 99% accuracy. Ascertainment of operative deaths occurring after discharge but within 30 days is more problematic, and this issue is currently the subject of an enhanced audit initiative.

Finally, the unit of inference for this study is the hospital. We did not examine surgeon–volume associations or the potential interaction between surgeon and hospital, both of which might have important ramifications.12,13

CONCLUSIONS

Although these findings demonstrate relatively stable performance over a wide range of CABG volumes, they do suggest the need to carefully scrutinize the performance of the lowest-volume hospitals. Although there are many excellent programs within this group, on average, their performance is the most problematic.12,13,16 Strategies for measuring and optimizing performance at such low-volume programs have been described.12,13

Efforts should continue to accurately measure risk-adjusted outcomes and adherence to existing best practices; to identify additional evidence-based, optimal care processes through registry-based research; and to apply such optimal care processes uniformly across programs of all sizes.

References

The model used for the composite score determination differs from the other models described in this study: (1) all 4 quality domains are analyzed jointly in a single model; (2) the model has a single coefficient for volume rather than a separate volume coefficient for each end point; and (3) the model was estimated by using hospital-aggregated summary data (to facilitate computation) instead of individual patient-level data. At each hospital, the number of ‘failures’ of each type (mortality, morbidity, ITA noncompliance, and medication noncompliance) was assumed to depend on volume, case mix (mortality and morbidity only), and a hospital-specific random-effects parameter. The inclusion of hospital-specific random effects is intended to capture variation in performance that is not explained by volume, case mix, or chance. The model is as follows:

\[
\logit \pi_j^{\text{morts}} = \alpha_{\text{morts}} + \beta_{\text{vol}} j + \gamma_{\text{meds}} \logit x_j^{\text{morts}} + \epsilon_j^{\text{morts}}
\]

\[
\logit \pi_j^{\text{mobs}} = \alpha_{\text{mobs}} + \beta_{\text{vol}} j + \gamma_{\text{meds}} \logit x_j^{\text{mobs}} + \epsilon_j^{\text{mobs}}
\]

\[
\logit \pi_j^{\text{mas}} = \alpha_{\text{mas}} + \beta_{\text{vol}} j + \epsilon_j^{\text{mas}}
\]

where \(j\) denotes the \(j\)th of 733 hospitals, \(\pi\) denotes the hospital’s true underlying failure rate for the indicated measure, \(\text{vol}\) denotes the hospital’s 2007 CABG volume, and \(x\) denotes a summary measure of case mix for the indicated outcome during the indicated time period.\(^4\) The remaining symbols represent unknown parameters to be estimated from the data. The overall composite score was defined as a weighted average of the domain-specific probabilities, with weights proportional to the reciprocal of the standard deviation. Average composite scores (posterior mean and 95% CI) were determined for each volume category, and the proportion of variation in the composite scores explained by volume was calculated.

**APPENDIX 1**

The model used for the composite score determination differs from the other models described in this study: (1) all 4 quality domains are analyzed jointly in a single model; (2) the model has a single coefficient for volume rather than a separate volume coefficient for each end point; and (3) the model was estimated by using hospital-aggregated summary data (to facilitate computation) instead of individual patient-level data. At each hospital, the number of ‘failures’ of each type (mortality, morbidity, ITA noncompliance, and medication noncompliance) was assumed to depend on volume, case mix (mortality and morbidity only), and a hospital-specific random-effects parameter. The inclusion of hospital-specific random effects is intended to capture variation in performance that is not explained by volume, case mix, or chance. The model is as follows:

\[
\logit \pi_j^{\text{morts}} = \alpha_{\text{morts}} + \beta_{\text{vol}} j + \gamma_{\text{meds}} \logit x_j^{\text{morts}} + \epsilon_j^{\text{morts}}
\]

\[
\logit \pi_j^{\text{mobs}} = \alpha_{\text{mobs}} + \beta_{\text{vol}} j + \gamma_{\text{meds}} \logit x_j^{\text{mobs}} + \epsilon_j^{\text{mobs}}
\]

\[
\logit \pi_j^{\text{mas}} = \alpha_{\text{mas}} + \beta_{\text{vol}} j + \epsilon_j^{\text{mas}}
\]

where \(j\) denotes the \(j\)th of 733 hospitals, \(\pi\) denotes the hospital’s true underlying failure rate for the indicated measure, \(\text{vol}\) denotes the hospital’s 2007 CABG volume, and \(x\) denotes a summary measure of case mix for the indicated outcome during the indicated time period.\(^4\) The remaining symbols represent unknown parameters to be estimated from the data. The overall composite score was defined as a weighted average of the domain-specific probabilities, with weights proportional to the reciprocal of the standard deviation. Average composite scores (posterior mean and 95% CI) were determined for each volume category, and the proportion of variation in the composite scores explained by volume was calculated.

**Discussion**

**Dr T. Bruce Ferguson** (Greenville, NC). Thank you, Dr Patterson. I have no disclosures. This study extends the CABG volume-outcome relationship beyond mortality alone to an evaluation of morbidity and processes of care. Among the most sophisticated of STS database analyses thus far, I congratulate Dr Shahian and his coworkers on this work. This analysis highlights what we all know: it is very difficult to provide continuous, highest-quality delivery of standardized care practices unless you ‘keep your edge.’ Although the magic number for annual CABG volume is unknown, re-examination in light of overall decreasing CABG volumes nationwide is merited. However, the lack of variability across the country is all the more remarkable when compared with the 10-fold differences in procedure volume for CABG. This is a direct testament to the deliberate focus on measuring and disseminating best practices by our specialty and in moving beyond risk-adjusted outcomes to processes of care. I have 2 questions for Dr Shahian.

First, the composite score domains—adjusted odds ratio shows differences in hospital mortality and ITA use but not in morbidity and medication failure. Because ITA use has been linked to improved short- and long-term mortality and because both ITA use and, to a degree, hospital mortality are surgeon dependent, does your analysis suggest that individual provider–level evaluation rather than hospital-level evaluation is inevitable? Second, STS risk-adjusted mortality and morbidity have been stable over the past 3 years and have not continued to decrease. The 2 lowest-volume groups comprise only 14.4% of the total patients undergoing CABG, and incremental improvement in their outcomes through whatever mechanism is not likely to affect these national benchmarks. What is next for CABG quality improvement? Thank you.

**Dr Shahian.** Thank you very much, Bruce. With regard to the first question about individual surgeon–versus hospital-level performance evaluation, our professional societies have historically focused on the latter. However, there are numerous reports from Ed Hannan in New York, as well as a paper presented 2 years ago at this meeting by Joe Carey from California, suggesting that the surgeon and his or her relationship with the rest of the team
are very important factors in achieving optimal outcomes. In fact, Bruce, you have mentioned one performance metric, use of the ITA, in which surgeons are completely in charge of an important process metric, and there are very few legitimate reasons not to use an ITA graft.

So how do we best incorporate surgeon performance into an overall quality-monitoring paradigm while hopefully not causing surgeons to become risk averse? Like politics, I think this might be best conducted locally, at least initially. For example, a hospital chief of cardiac surgery or the leader of a practice group would carefully review his or her performance reports from the STS database, notice a low score resulting from suboptimal ITA use, drill down internally to the surgeon level, and then discuss this with the appropriate staff. As a profession, we have been rather reluctant to have those kinds of difficult conversations. However, if we do not accept this responsibility, then others will force the issue, and the result will be even more onerous.

Your second question involved where we go from here in quality improvement. First, I would not discount the value of focusing some improvement efforts on the 2 lowest-volume hospital categories. Even though they cared for only 14% of the patients in our series, they include about 36% of hospitals overall and their performance in most areas appears inferior to that of larger hospitals.

So what do I think we need to be doing in the quality assurance and improvement areas? First, we need to develop a broader portfolio of evidence-based processes of care that is much more comprehensive than we currently measure. We need to implement these nationally using state and regional peer collaboration, and we need to reduce the variability that currently exists for many of such measures. We also need enhanced performance monitoring, particularly in the low-volume programs, where this is so problematic. I personally think that tertiary sponsorship or oversight of the lowest-volume programs is very important, and this is a model that has worked well in Massachusetts. Finally, I think there are some extremely low-volume programs for which regionalization would be a better option. Thank you.

Dr Craig R. Smith (New York, NY). Very simply, did you look at geography as a variable at all, state or region, or, more interestingly, is rural low volume the same as urban low volume?

Dr Shahian. We did not look at that, but it would be an excellent next step.

Dr John D. Puskas (Atlanta, Ga). David, I enjoyed your presentation. Did you have a chance to look at the effect of low- or high-volume centers on mortality and morbidity after off-pump compared with on-pump surgery? We are still in an era in which many consider that a more technically demanding operation. Does an institution or surgeon need to be doing more of it to hit a decent benchmark in quality?

Dr Shahian. John, we did not do that, but it would be an excellent follow-on study. You are absolutely correct that it is a very interesting and important area to investigate.

Dr Marc R. Moon (St Louis, Mo). You said that the ITA use was lowest in the low-volume centers but also that the acuity was the highest. Are you sure there was not a relationship between lack of ITA use and patients being sicker when they went for surgical intervention?

Dr Shahian. Well, the trend adjusted for acuity still showed the same finding. Granted, the absolute differences are small. For all these differences, many of them are statistically significant because we have such a large sample size, but there was about a 1% or 2% difference, and the lowest use among all programs—and some of them had very low use rates—were at the low-volume end. We had some programs with 70% or 60% ITA use, and they are all at the low end of the volume spectrum. I do not think that is completely accounted for by acuity or other factors.