

The Society of Thoracic Surgeons Composite Score for Rating Program Performance for Lobectomy for Lung Cancer

Benjamin D. Kozower, MD, MPH, Sean M. O'Brien, PhD, Andrzej S. Kosinski, PhD, Mitchell J. Magee, MD, Rachel Dokholyan, MPH, Jeffery P. Jacobs, MD, David M. Shahian, MD, Cameron D. Wright, MD, and Felix G. Fernandez, MD

University of Virginia, Charlottesville, Virginia; Duke Clinical Research Institute, Durham, North Carolina; Medical City Dallas Hospital, Dallas, Texas; All Children's Hospital, John Hopkins University, St. Petersburg, Florida; Massachusetts General Hospital, Boston, Massachusetts; and Emory University, Atlanta, Georgia

Background. The Society of Thoracic Surgeons (STS) has developed multidimensional composite quality measures for common cardiac surgery procedures. This first composite measure for general thoracic surgery evaluates STS participant performance for lobectomy in lung cancer patients.

Methods. The STS lobectomy composite score is composed of two outcomes: risk-adjusted mortality; and any-or-none, risk-adjusted major complications. General Thoracic Surgery Database data were included from 2011 to 2014 to provide adequate sample size, and 95% Bayesian credible intervals were used to determine "star ratings." The STS participants were also compared with national benchmarks (including non-STs participants) using the National Inpatient Sample. Comparisons of discharge mortality, postoperative length of stay, and percent of stage I lung cancers resected using minimally invasive approaches are not included in star ratings but will be reported to participants in STS feedback reports.

Results. The study population included 20,657 lobectomy patients from 231 participating centers. Operative mortality was 1.5%, major complication rate was 9.6%, and median postoperative length of stay was 4 days. Risk-adjusted mortality and major complication rates varied threefold from highest performing (three-star) to lowest performing (one-star) programs. Approximately 5% of participants were one-star, 7% were three-star, and 88% were two-star programs.

Conclusions. The STS has developed the first general thoracic surgery quality composite measure to compare programs performing lobectomy for lung cancer. This measure will be used for quality assessment and provider feedback, and will be made available for voluntary public reporting.

(Ann Thorac Surg 2016;101:1379–87)

© 2016 by The Society of Thoracic Surgeons

The Society of Thoracic Surgeons (STS) Quality Measurement Task Force has developed three composite performance measures for common procedures in adult cardiac surgery [1–3]. These measures are available for public reporting on the STS and Consumer Reports websites, and participation is voluntary [4]. The STS leadership believes that the public has a right to see and understand the quality of surgical outcomes, and regards public reporting as an ethical responsibility. High-quality clinical data, procedure-specific risk-adjustment models, and multidimensional composite measures have provided the foundation for this STS public reporting initiative.

The STS General Thoracic Surgery Database (GTSD) has a risk-adjustment model for morbidity and mortality after lung cancer resection [5]. Because lobectomy for cancer is the most commonly performed major general thoracic surgical procedure in the GTSD, the working group selected this procedure for its first thoracic quality measure. The lobectomy composite was designed as a two-domain measure including risk-adjusted mortality and major complications. This methodology mirrors the development of the STS aortic valve replacement (AVR) and AVR plus coronary artery bypass graft (CABG) measures.

This report describes the development of the first composite measure for general thoracic surgery evaluating participant performance for mortality and major complications after lobectomy for lung cancer. As

Accepted for publication Oct 26, 2015.

Presented at the Fifty-first Annual Meeting of The Society of Thoracic Surgeons, San Diego, CA, Jan 24–28, 2015.

Winner of the Richard E. Clark Award for General Thoracic Surgery.

Address correspondence to Dr Kozower, University of Virginia, Box 800679, Charlottesville, VA 22908-0679; email: bdk8g@virginia.edu.

The Appendix can be viewed in the online version of this article [<http://dx.doi.org/10.1016/j.athoracsur.2015.10.081>] on <http://www.annalsthoracicsurgery.org>.

additional measures of performance, we also compare STS participants with national lobectomy outcomes data from the National Inpatient Sample (NIS) and report the utilization of minimally invasive lobectomy by STS participants.

Material and Methods

Study Cohort

The STS GTSD was queried for all patients treated with lobectomy for lung cancer between July 1, 2011, and June 30, 2014 (all records were in Data Collection Form v2.081 and v2.2). We selected a time frame of 3 years because it provided an adequate sample size to perform the analyses while recognizing that the inclusion of data from more remote time frames might make scores less relevant to current practice. That is consistent with current GTSD practice using a 3-year time frame for lung and esophageal cancer models. We excluded patients with nonelective status, occult or stage 0 tumors, American Society of Anesthesiologists class VI, and with missing data for age, sex, or discharge mortality status. The final study population was 20,657 operations from 231 participating centers.

Outcome Definitions

Postoperative events were defined by the STS GTSD guidelines [6]. Operative mortality is defined as death during the same hospitalization as surgery or within 30 days of the procedure [7]. Major complications are defined as one or more of the following: pneumonia, acute respiratory distress syndrome, bronchopleural fistula, pulmonary embolus, initial ventilator support greater than 48 hours, reintubation/respiratory failure, tracheostomy, myocardial infarction, or unexpected return to the operating room. The previous definition of major complications after lung cancer resection included reoperation for bleeding rather than unexpected return to the operating room for any cause [5].

Estimation of Risk-Adjusted Mortality and Complication Rates

The STS lobectomy composite score is a combination of two risk-adjusted outcome metrics: operative mortality and major complications. Participant-specific risk-adjusted rates of these endpoints were estimated in a Bayesian hierarchical model, as detailed in the [Appendix](#). Covariates in this model were taken from the previous lung cancer resection model [5]: age, sex, year of operation, body mass index, hypertension, steroid therapy, congestive heart failure, coronary artery disease, peripheral vascular disease, reoperation, preoperative chemotherapy within 6 months, cerebrovascular disease, diabetes mellitus, renal failure, dialysis, past smoker, current smoker, forced expiratory volume in 1 second percent of predicted, Zubrod score (linear plus quadratic), American Society of Anesthesiologists class (linear plus quadratic), and pathologic stage as defined by the American Joint Committee on Cancer cancer staging manual, 6th edition.

Estimation of Composite Scores and Star Ratings

The composite score was calculated as a weighted sum of (1 minus the risk-adjusted mortality rate) and (1 minus the risk-adjusted complication rate). Mortality and major complications were weighted inversely by their respective standard deviations across participants. This procedure is equivalent to first rescaling mortality and complications by their respective standard deviations and then assigning equal weighting to the rescaled mortality rate and rescaled complication rate. This is the same methodology used for other STS composite measures [3, 8]. As described in the [Appendix](#), composite scores were estimated for each STS participant and reported with 95% Bayesian credible interval (CrI).

To assign STS participants into performance categories, each participant's composite score was compared with what would have been expected from an average provider in the STS GTSD with a similar case-mix. Participants were classified as having a lower-than-expected composite score (one star) if their 95% CrI fell entirely below the STS average composite score, as having a higher-than-expected composite score (three star) if their 95% CrI fell entirely above the STS average score, and as having a composite score not statistically distinguishable from expected (two star) if their 95% CrI overlapped the STS average composite score. In addition to calculating star ratings based on 95% CrI, we also explored the use of 80%, 90%, and 98% CrI for this purpose. The objective was to identify the CrI with the highest level of certainty, while also producing a reasonable degree of discrimination among providers. Results of previous STS adult cardiac composite measures have been presented with both numerical scores (point estimates with confidence intervals) and star ratings. Star ratings are provided as an aid to interpretation by patients, as many studies have shown that most patients cannot accurately interpret and use numerical data to make informed decisions [9]. Therefore, the STS reports the star ratings along with the detailed, risk-adjusted outcomes in an effort to best serve different stakeholders [4].

Reliability Estimation

A Bayesian estimate of reliability was calculated and reported with 95% CrI. (See the [Appendix](#) for detailed methods [2].) Reliability is a key metric of the suitability of a measure for profiling because it describes how well one can confidently distinguish the performance of one provider from another [10]. Conceptually, reliability is the ratio of signal to noise. The signal is the proportion of the variability in measured performance that can be explained by real differences in performance. The noise is the random statistical variation that occurs in routine clinical practice. A reliability of zero implies that all the variability in a measure is attributable to measurement error. A reliability of one implies that all the variability is attributable to real differences in performance. There are three main drivers of reliability: sample size, true performance differences between providers, and measurement error [10].

Additional Outcome and Process Measures for Biannual Reports

Unlike adult cardiac surgery, in which almost all procedures in the United States are performed by STS database participants, many general thoracic procedures including lobectomy for cancer are performed by non-GTSD members, both general surgeons and, primarily, cardiac surgeons. Because lung cancer resection outcomes for STS GTSD participants are better than national averages [11], it is also important to compare the performance of these programs with those of all US hospitals performing lobectomy, not just those in the GTSD. For example, a hospital with a two-star STS rating,

Table 1. Characteristics of Study Population

Characteristics	Values
Age	67 ± 10.2
Male	45.3%
Body mass index, kg/m ²	28 ± 6.2
Hypertension	62.2%
Steroid use	3.4%
Congestive heart failure	2.9%
Coronary artery disease	22.0%
Peripheral arterial disease	9.1%
Reoperation	5.6%
Preoperative chemotherapy	6.3%
Cerebrovascular disease	8.2%
Diabetes mellitus	18.8%
Renal failure	1.2%
Dialysis	0.6%
Past smoker	61.2%
Current smoker	23.7%
FEV ₁ percent of predicted	83 ± 20.2
<60	17.0%
60–99	67.4%
100+	20.2%
Zubrod score	
0	43.1%
1	52.9%
2	3.3%
3	0.6%
4	0.1%
5	<0.1%
ASA class	
I	0.3%
II	15.4%
III	75.0%
IV	9.2%
V	<0.1%
Stage	
I	69.2%
II	18.3%
III	10.8%
IV	1.6%

ASA = American Society of Anesthesiologists; FEV₁ = forced expiratory volume in 1 second.

Table 2. Number of Hospitals, Lobectomies, and Endpoint Events

Variables	Values
Number of participants	231
Number of lobectomy records	20,657
Frequency of operative mortality	303 (1.5%)
Frequency of major complications	1984 (9.6%)
Pneumonia	884 (4.3%)
Acute respiratory distress syndrome	136 (0.7%)
Bronchopleural fistula	90 (0.4%)
Pulmonary embolus	110 (0.5%)
Initial ventilator support >48 hours	94 (0.5%)
Reintubation	697 (3.4%)
Tracheostomy	202 (1.0%)
Myocardial infarction	76 (0.4%)
Unexpected return to operating room	806 (3.9%)
Length of stay, median days (IQR)	4 (3–7)

IQR = interquartile range.

performing at an expected level compared only to other STS GTSD participants, might be a better than average performer when compared with national outcomes from all hospitals performing lobectomies for lung cancer in the United States. Similarly, an STS GTSD participating hospital with a one-star STS rating, performing below the expected level in the STS database, could be an average performer if compared with all US hospitals performing lobectomies. Therefore, on STS biannual feedback reports and on future public reporting websites, STS participant scores will be reported and compared with two separate benchmarks: (1) other STS GTSD participants, and (2) national outcomes for all programs performing lobectomies. Discharge mortality and postoperative length of stay will be directly compared between the STS and NIS. Because these databases use different comorbidity definitions that are not directly comparable, comparisons between them are not risk adjusted.

We selected the NIS as the best available source for overall national outcomes. The NIS is the largest, all-payer inpatient database available in the United States, representing a 20% sample of all hospital discharges from nonfederal facilities [12]. This database is maintained by the Agency for Healthcare Research and Quality. It includes almost 8 million inpatient hospital discharge abstracts collected annually for patients of all ages and all sources of insurance. Each discharge record includes a weight that represents the relative proportion of the total US hospital patient population accounted for by that record [13]. This dataset is broadly

Table 3. Composite Score Weights

	Mortality	Major Complications
Standard deviation	0.84	3.42
Weight	0.80	0.20

Table 4. Reliability (Proportion of Signal to Noise Variation) of Star Ratings Based on Program Volume Thresholds

Variable	No Minimum	≥30 Cases	≥50 Cases	≥100 Cases	≥150 Cases
No. of participants	231	172	136	80	40
Reliability (95% CrI)	46% (35%–55%)	56% (45%–66%)	61% (50%–70%)	68% (56%–79%)	74% (59%–85%)

CrI = credible interval.

representative of persons in the US population who were hospitalized that year and contains the most generalizable data to represent national lung cancer resection outcomes.

Minimally invasive approaches (pure video-assisted thoracic surgery or robotic) for lobectomy are now commonplace among STS surgeons and are performed with increasing frequency every year. Although no randomized clinical trial has been performed to examine the perioperative or long-term outcomes, STS data, administrative data, and numerous case series report a reduction in perioperative morbidity and equivalent long-term survival when minimally invasive approaches are used instead of a standard thoracotomy [5, 14, 15]. Specifically, we have previously demonstrated that minimally invasive lung cancer resection has a 50% reduction in major complications compared with a thoracotomy approach, adjusted for age, sex, and comorbidities [5]. Most recently, a propensity matched analysis of prospectively collected Cancer and Leukemia Group B (CALGB) 140202 data demonstrated that minimally invasive lobectomy had a shorter hospital length of stay, fewer complications, and greater likelihood of independent discharge to home compared with open lobectomy for early stage lung cancer [16]. Some surgeons use a minimally invasive approach for all lobectomies regardless of stage, but many surgeons are reluctant to limit access for higher stage tumors. However, there is a general consensus among STS surgeons and the STS GTSD task force that stage I lung cancer is usually resectable with a minimally invasive approach. Because many patients desire a minimally invasive approach, and our STS data and other published data demonstrate improved risk-adjusted outcomes, the working group considered it appropriate to include the percent of minimally invasive lobectomies for stage I lung cancer as a process measure on the STS biannual reports and on future public reporting efforts, but not incorporate it into the composite measure and star ratings.

Results

The characteristics of the study population are shown in Table 1. The number of procedures, hospitals, and endpoints for mortality and major complications are shown in Table 2. The 3-year sample provided patient cohorts and numbers of endpoints that were similar to previous thoracic surgery modeling efforts and consistent with the quality measures developed for adult cardiac surgery [2, 3].

Because of the rescaling of its mortality and complication components based on the reciprocals of their standard deviations, these two components of the overall composite score are effectively weighted differently, as shown in Table 3. Mortality is weighted approximately four times that of a major complication in the quality measure, consistent with the adult cardiac surgery quality measures [3]. The GTSD working group believes this is an improvement from its previous lung cancer resection model in which mortality and major morbidity were weighted equally.

Table 4 illustrates the reliability for the lobectomy star ratings based on volume thresholds. Using the threshold volume of performing 30 lobectomies over the 3-year study period, the reliability of the STS lobectomy composite measure was 0.56 (95% CrI: 0.45 to 0.66), similar to the reliability of the STS AVR plus CABG measure, which was 0.51 (95% CrI: 0.46 to 0.55) [3]. Therefore, only programs performing a minimum of 30 lobectomies for lung cancer will be eligible for a star rating as there is insufficient information about the 25% of programs performing fewer than 10 lobectomies per year to provide them with a reliable star rating.

Table 5 shows the number of STS participants classified as high- (three-star), average- (two-star), or low- (one-star) performing centers using 80%, 90%, 95%, and 98% Bayesian CrI. Although there is no clear statistical criterion on which to base our choice of the most appropriate CrI, from a practical perspective we believe that 98% CrI

Table 5. Number of STS Participants Identified as Different From STS Average Based on Different Probability Thresholds^a

Criterion for Categorizing Participants	Worse Than STS Average (One Star)	Indistinguishable From STS Average (Two Star)	Better Than STS Average (Three Star)
98% CrI falls above/below STS average	5 (2.9%)	160 (93.0%)	7 (4.1%)
95% CrI falls above/below STS average	8 (4.7%)	152 (88.4%)	12 (7.0%)
90% CrI falls above/below STS average	12 (7.0%)	142 (82.6%)	18 (10.5%)
80% CrI falls above/below STS average	16 (9.3%)	128 (74.4%)	28 (16.3%)

^a Among 172 hospitals with at least 30 lobectomies.

CrI = credible interval; STS = The Society of Thoracic Surgeons.

Table 6. Construct Validity: Mortality and Major Complication Rates Vary Across Star Ratings^a

Variable	One Star	Two Star	Three Star	All Programs
Operative mortality (95% CrI)	3.2% (1.6%–5.9%)	1.6% (0.6%–3.5%)	0.9% (0.4%–1.6%)	1.7% (0.6%–3.9%)
Major complication (95% CrI)	17.1% (11.3%–24.2%)	10.1% (5.1%–16.9%)	6.5% (3.7%–9.6%)	10.2% (4.8%–18.4%)

^a Among 172 hospitals with at least 30 lobectomies.

CrI = credible interval.

provided higher specificity but inadequate differentiation among programs. Conversely, although the percentage of high- and low-performing programs was substantially larger with 80% CrI, there was insufficient certainty about the participant's classification to assure face validity. Based on these practical considerations and what the STS has done previously for their quality measures, we chose 95% CrI (corresponding to 97.5% Bayesian probability). This yielded 4.7% one- star programs (8 of 172) and 7.0% three-star programs (12 of 172), or nearly 11% overall high-performing or low-performing centers. This seemed an appropriate compromise between providing acceptable probability of accurate classification along with reasonable discrimination among programs. Table 6 demonstrates construct validity of the star ratings as the mortality and major complication rates decrease monotonically from one-star (below average) to three-star (above average) participants.

Figure 1 illustrates participating programs sorted in order of increasing composite score. The average composite score was 96.8. Participants to the right of the figure, with higher composite scores and credible intervals that do not cross the average, are high-performing (three-star) programs with fewer mortalities and major complications.

The unadjusted comparisons between the STS and NIS databases are shown in Table 7. As previously demonstrated, participants in the STS GTSD have lower discharge mortality and postoperative length of stay after lobectomy than national benchmarks. The percentage of minimally invasive lobectomies for stage I lung cancer

in the STS GTSD was 63.7% (10,112 of 15,869; 95% confidence interval: 63.0% to 64.5%). Figure 2 demonstrates the distribution of the percentage of minimally invasive lobectomies for stage I lung cancer performed by STS participating centers. As shown in the figure, there is considerable variation in the use of minimally invasive techniques. Individual program performance on all three domains (star rating, participant versus NIS comparison, and percent of minimally invasive lobectomies) will be reported to participants in biannual reports.

Comment

We have described the development and operational characteristics of the first composite quality measure for general thoracic surgery in the STS GTSD. The lobectomy composite includes both risk-adjusted mortality and risk-adjusted, any-or-none major complications, but no process measures. This composite measure adds considerable value to members of the GTSD by accurately comparing participants to their STS peers, facilitating quality assessment with biannual reports, and providing a comparison of STS participants to national benchmarks.

The use of minimally invasive techniques (video-assisted thoracic surgery or robotic) was considered as a process measure because it is associated with reduced morbidity and length of stay and because the public and stakeholders are increasingly interested in this approach [5, 14]. However, it is not a National Quality Forum-endorsed quality metric, and there is a general move away from process measures toward a focus on outcome

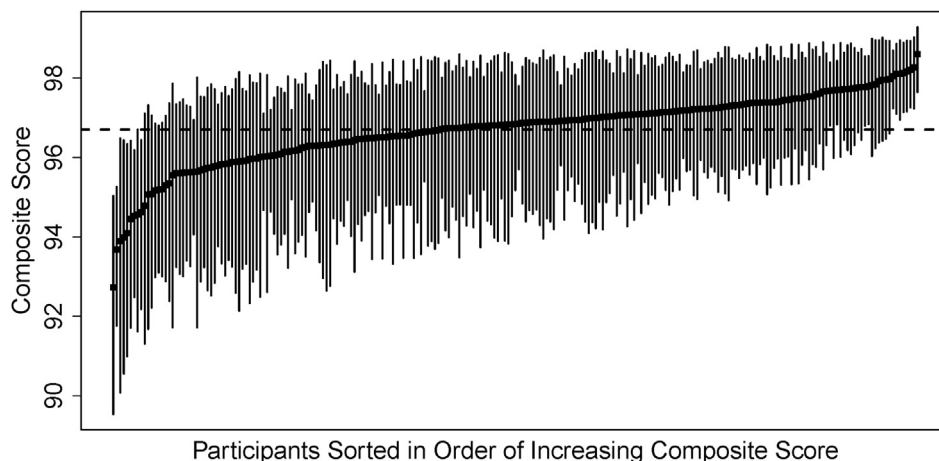


Fig 1. Distribution of participant's composite score for lobectomy. Participating programs are sorted in order of increasing composite score. The average composite score was 96.8. Participants to the right on the figure, with higher composite scores and credible intervals that do not cross the average, are high-performing (three-star) programs with fewer mortalities and major complications.

Table 7. Comparison Between The Society of Thoracic Surgeons and National Inpatient Sample Risk-Unadjusted Outcomes for Lobectomy

Variable	STS	NIS
Discharge mortality	n = 20,777	n = 26,015
Percent	1.01% (209/20,777)	1.77% (460/26,015)
95% CI	0.87%–1.15%	(1.39%–2.14%)
Postoperative LOS, days	n = 20,753	n = 26,015
Mean, 95% CI	5.9 (5.8–6.0)	7.3 (7.1–7.5)
Median, IQR	4.0 (3.0–7.0)	5.3 (3.5–7.8)

CI = confidence interval; IQR = interquartile range; LOS = length of stay; NIS = National Inpatient Sample; STS = The Society of Thoracic Surgeons.

measures. Therefore, we will report the percentage of minimally invasive lobectomies for stage I lung cancer performed by a program on the STS biannual reports but have elected not to include it in the composite measure for lobectomy.

The lobectomy composite has several limitations. First, participation in the STS GTSD is voluntary and represents fewer than 50% of lung cancer resections performed nationally [11]. Importantly, independent and external audits have verified that the STS GTSD has a 95% data accuracy rate and that programs have submitted consistently accurate data [17]. However, the STS GTSD has less national penetration than the STS Adult Cardiac Surgery Database, which includes more than 90% of procedures performed in the country.

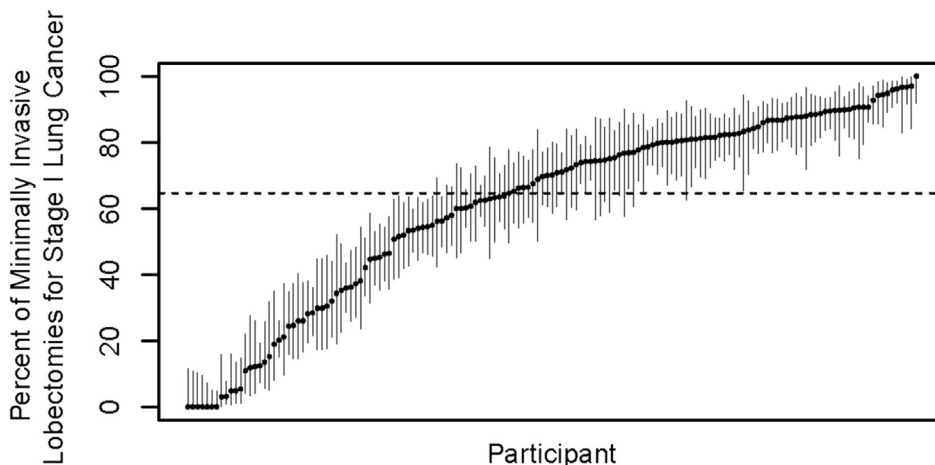
Second, STS GTSD participants have lower mortality and length of stay than national benchmarks [11]. Therefore, STS results are not generalizable to all hospitals in the United States that perform lobectomy for cancer. For this reason, we have included an unadjusted comparison of STS participants to more broadly representative national lobectomy outcomes, using the NIS. By publically recognizing and reporting the improved outcomes among GTSD participants compared with national

benchmarks, we hope that overall participation in the GTSD will increase, outcomes will continue to improve, and participants will choose to participate in the voluntary public reporting option. The STS is planning on making this measure available on the STS and Consumer Reports public reporting websites in 2017 after a 1-year pilot period for participant feedback.

Third, the reliability of the lobectomy composite metric including all programs is 0.46, which is a lower signal-to-noise ratio than reported for some of the adult cardiac surgery quality measures, such as 0.51 for the AVR plus CABG composite score. This is in part explained by the overall smaller case volumes per hospital and the presence of several hospitals with only a few cases. In hospitals with at least 30 cases over 3 years (a more stringent requirement to insure adequate sample size), the estimated signal-to-noise reliability was 0.56, comparable to the STS composite score for AVR plus CABG. Therefore, the composite rating will only be applied to programs performing at least 30 lobectomies for lung cancer over the study period. Unfortunately, this means that the star rating will only be applied to 172 of the 231 programs. Not rating 59 programs, or 25% of participating centers, is significant but given their very low procedure volume, we simply do not know enough about their true performance to reliably assign them a star rating.

Traditionally, a reliability of 0.7 has been considered “sufficient” for performance measures [10]. However, health care measures have often had lower levels of reliability owing to relatively small sample sizes, and measures with lower reliabilities have been implemented for quality measurement. For example, a recent standardized readmission ratio for dialysis facilities developed by the Centers for Medicare and Medicaid has a reliability across 4 calendar years ranging between 0.49 and 0.54 [18]. Recent work by Dimick and colleagues [19] demonstrated that surgical quality metrics rarely meet a reliability bar of 0.5, largely because of sample size issues. As an important adjunct to their work, the same group investigated whether measures

Fig 2. Distribution of participant's use of minimally invasive resection for stage I lung cancer, demonstrating significant variation in the use of minimally invasive resection (video-assisted thoracic surgery or robotic) for stage I lung cancer among participants of The Society of Thoracic Surgeons General Thoracic Surgery Database.



with progressively lower reliability were useful in predicting future performance. The investigators demonstrated that even at very low reliability (levels of 0.1 and 0.2), risk-adjusted outcome measures can distinguish “best” and “worst” hospitals’ surgical performance [20]. They concluded that in the context of selective referral/public reporting, commonly accepted reliability thresholds are too high.

Therefore, we believe that this lobectomy quality measure, with a reliability of 0.56, is appropriate for quality measurement and for public reporting. Our STS data are the best available clinical data, and this lobectomy measure is the most accurate model available for thoracic surgery. Developing quality composite measures is an iterative process, and future measures, including larger sample sizes and longer term outcomes, are likely to improve our ability to measure differences in performance.

The lobectomy composite measure is composed of two short-term outcomes: perioperative mortality and major complications. Long-term outcomes are extremely important, particularly for a disease such as lung cancer when adequacy of staging and resection impact long-term survival. At present, these longer term outcomes are not available in the STS GTSD. Accordingly, the GTSD has updated its data specifications to include 5-year survival beginning with the 2015 data harvest. In addition, the STS established a linkage of their data to administrative data from the Centers for Medicare and Medicaid Services for CABG [21]. Plans to extend this linkage to lung cancer resections will also provide longitudinal follow-up for patients aged 65 years and more. These exciting developments within the STS database will facilitate exploration of including longer term outcomes such as 90-day mortality, 1-year mortality, or even 5-year mortality. For example, it has been shown in several databases that mortality after lung cancer resection doubles between 30 and 90 days [22, 23]. Including longer outcomes will be extremely important for refinement of this and future general thoracic quality measures. They are important to patients and other stakeholders, they will increase the number of outcome events, and they will further increase the ability of quality measures to differentiate outliers.

The lobectomy composite identifies nearly 12% of programs as statistically above or below average performance (4.7% below average, 7.0% above average). The majority of programs, 88.4% (152 of 172), are classified as expected or average performers. Although the percentage of above or below average performers could be increased by using 90% CrI or 80% CrI, the cost of this increased sensitivity would be a loss in specificity, with more false positive classifications. The GTSD working group believes that, for our data, a 95% CrI provides the best balance of sensitivity and specificity, and this approach was similar to that used in previous STS composite measures [2, 3]. As described above, the inclusion of longer term outcomes in the future is likely to improve the measures ability to accurately differentiate participant performance.

In conclusion, the STS has developed a two-domain composite performance measure for lobectomy for lung cancer based on risk-adjusted mortality and the risk-adjusted occurrence of any of eight major complications. This composite measure identifies nearly 12% of participating programs as outliers. Three other endpoints will be reported on the STS biannual reports and be available for the future voluntary public reporting effort. Two of these are participant outcome comparisons for discharge mortality and postoperative length of stay, comparing STS programs with a sample of all US hospitals that perform lobectomy for cancer. The third is a process measure for the use of minimally invasive lobectomy, either video-assisted thoracic surgery or robotic, for stage I lung cancers. That will compare participants with their peers in the STS.

The authors are very grateful for the hard work and dedication of Jane Han, MSW, and Donna McDonald, MPH, RN, from The Society of Thoracic Surgeons, Chicago, Illinois.

References

1. Shahian DM, O'Brien SM, Filardo G, et al. The Society of Thoracic Surgeons 2008 cardiac surgery risk models: part 1—coronary artery bypass grafting surgery. *Ann Thorac Surg* 2009;88(Suppl):2–22.
2. 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;94:2166–71.
3. Shahian DM, He X, Jacobs JP, et al. The STS AVR plus CABG composite score: a report of the STS Quality Measurement Task Force. *Ann Thorac Surg* 2014;97:1604–9.
4. The Society of Thoracic Surgeons. Public reporting. Available at <http://www.sts.org/quality-research-patient-safety/sts-public-reporting-online>. Accessed December 26, 2014.
5. Kozower BD, Sheng S, O'Brien SM, et al. STS database risk models: predictors of mortality and major morbidity for lung cancer resection. *Ann Thorac Surg* 2010;90:875–83.
6. STS general thoracic surgery database: database collection. Available at <http://www.sts.org/quality-research-patient-safety/national-database/database-managers/general-thoracic-surgery-databa-1>. Accessed September 21, 2015.
7. Overman DM, Jacobs JP, Prager RL, et al. Report from The Society of Thoracic Surgeons National Database Workforce: clarifying the definition of operative mortality. *World J Pediatr Congenit Heart Surg* 2013;4:10–2.
8. Rankin JS, He X, O'Brien SM, et al. The Society of Thoracic Surgeons risk model for operative mortality after multiple valve surgery. *Ann Thorac Surg* 2013;95:1484–90.
9. Hibbard JH, Peters E, Slovic P, Finucane ML, Tusler M. Making health care quality reports easier to use. *Jt Comm J Qual Improv* 2001;27:591–604.
10. Adams J. The reliability of provider profiling: a tutorial. 2009. Available at http://www.rand.org/pubs/technical_reports/tr653.html. Accessed September 21, 2015.
11. LaPar DJ, Bhamidipati CM, Lau CL, Jones DR, Kozower BD. The Society of Thoracic Surgeons general thoracic surgery database: establishing generalizability to national lung cancer resection outcomes. *Ann Thorac Surg* 2012;94:216–21.
12. Davies HE, Davies RJO, Davies CWH, for the BTS Plural Disease Guideline Group. Management of pleural infection in adults: British Thoracic Society pleural disease guideline 2010. *Thorax* 2010;65(Suppl 2):ii41–53.
13. Whalen D, Houchens R, Elixhauser A. 2004 HCUP Nationwide Inpatient Sample (NIS) comparison report. HCUP methods series report # 2007-03. Online December 2, 2007.

- US Agency for Healthcare Research and Quality. Available at <http://www.hcup-us.ahrq.gov/db/nation/nis/nisdbdocumentation.jsp>. Accessed December 13, 2015.
14. Swanson SJ, Meyers BF, Gunnarsson CL, et al. Video-assisted thoracoscopic lobectomy is less costly and morbid than open lobectomy: a retrospective multiinstitutional database analysis. *Ann Thorac Surg* 2012;93:1027–32.
 15. Farjah F, Backhus LM, Varghese TK, et al. Ninety-day costs of video-assisted thoracic surgery versus open lobectomy for lung cancer. *Ann Thorac Surg* 2014;98:191–6.
 16. Nwogu CE, D'Cunha J, Pang H, et al. VATS lobectomy has better perioperative outcomes than open lobectomy: CALGB 31001, an ancillary analysis of CALGB 140202 (Alliance). *Ann Thorac Surg* 2015;99:399–405.
 17. Magee MJ, Wright CD, McDonald D, Fernandez FG, Kozower BD. External validation of The Society of Thoracic Surgeons general thoracic surgery database. *Ann Thorac Surg* 2013;96:1734–9.
 18. Centers for Medicare & Medicaid Services. Standardized readmission ratio for dialysis facilities. 2014. Available at <https://www.cms.gov/outreach-and-education/outreach/npc/national-provider-calls-and-events-items/2014-04-17-esrd-call.html>. Accessed September 30, 2015.
 19. Krell RW, Hozain A, Kao LS, Dimick JB. Reliability of risk-adjusted outcomes for profiling hospital surgical quality. *JAMA Surg* 2014;149:467–74.
 20. Krell RW, Staiger DO, Dimick JB. Reliability of surgical outcomes for predicting future hospital performance. *Med Care* 2014;52:565–71.
 21. Jacobs JP, Edwards FH, Shahian DM, et al. Successful linking of The Society of Thoracic Surgeons adult cardiac surgery database to Centers for Medicare and Medicaid Services Medicare data. *Ann Thorac Surg* 2010;90:1150–7.
 22. Pezzi CM, Mallin K, Mendez AS, Greer Gay E, Putnam JB. Ninety-day mortality after resection for lung cancer is nearly double 30-day mortality. *J Thorac Cardiovasc Surg* 2014;148:2269–78.
 23. Hu Y, McMurry TL, Wells KM, Isbell JM, Stukenborg GJ, Kozower BD. Postoperative mortality is an inadequate quality indicator for lung cancer resection. *Ann Thorac Surg* 2014;97:973–9.

DISCUSSION

DR FARHOOD FARJAH (Seattle, WA): Thank you for the opportunity to discuss this work. Dr Kozower, congratulations. I really enjoyed your presentation and thank you for sending me the manuscript in advance.

You and your team set out to develop a composite performance measure to evaluate the quality of care delivered to lung cancer patients undergoing lobectomy. Your work complements other recent efforts by The Society of Thoracic Surgeons (STS) Workforce on National Databases to develop composite performance measures for adult cardiac surgery. The overall intent here is to improve quality and facilitate transparency using high-quality clinical data and state-of-the-art statistical methodology.

You used empirical methods that gave death a weight four times greater than complications. From a surgeon's perspective, this distribution is reasonable. After all, we have historically measured safety by measuring operative mortality rates. But one of the several advantages of participating in a clinical database is the ability to robustly ascertain postoperative adverse events. I wonder if an empirical approach to weighting postoperative outcomes underestimates the impact of complications on patients. Patients may perceive some complications to approach, equal, or even be worse than death, and if so and if individual event rates vary from one institution to another, then performance rankings may change. So here are my questions.

Are there any examples of composite performance measures that apply patient-derived weights to individual adverse events? And if not, do you think it is possible to do so in a methodologically rigorous fashion? And do you think quality improvement initiatives would benefit from engaging and collaborating with patients in the design and implementation of performance metrics? Thank you.

DR KOZOWER: Thank you for your kind comments, Dr Farjah, and also your very insightful questions. I am going to answer the third one first.

I agree that a more patient-centered approach would have a lot to offer, particularly in 2015 with the development of the Patient Centered Outcomes Research Institute. I think it would be a wonderful thing for the STS to consider for inclusion in their database and subsequent risk adjustment models.

Regarding your first question, I am not aware of any well-done composite measure that includes patient-derived utilities. I do

think it is possible, although it would not be easy to do. As we have seen with the incorporation of patient reported outcomes, it can be done in a research setting but is harder to do in routine clinical practice. I think we are both aware of work by Cykert, which found that for early stage lung cancer as many as 30% of stage I lung cancer patients do not undergo surgical resection because of their fears for postoperative complications. Importantly, you appropriately point out that some complications may indeed be worse than death. For example, how would a patient weight their outcome if they had a tracheostomy after lobectomy with a prolonged hospital stay, end up in a skilled nursing facility and ultimately die or never return to their baseline performance status?

I do think the good news for us is that you are now part of the STS general thoracic database task force, and this is something that we can look at in the future.

DR MICHAEL T. JAKLITSCH (Boston, MA): I also want to comment about the morbidity component of this scale, and the way it seems to be presented, it perhaps is just a little bit too crude. I am encouraged by the idea of grading morbidity on a scale of 1 to 4 and the value of this, and I think that the aggressive looking for morbidity is what drives mortality down.

In the short term that I have been a thoracic surgeon, the operative mortality for lobectomy has dropped from 3.5% to 1.2%. The more we find the grade 1 and 2 morbidities, the more we prevent them from becoming grade 5 morbidities and death.

So a crude system that is just having you report morbidity as a factor hurts the big centers. If I go to a community hospital, they could not find a morbidity if you pointed it out to them. Whereas the aggressive academic centers, they are doing bronchoscopy for atelectasis and preventing that from going up to a grade 3. They are doing surveillance ultrasonography to find those blood clots when they are grade 1 and preventing them from getting to grade 3. So we need a system that incorporates the grade 1 to 4 morbidity grids.

DR KOZOWER: Dr Jaklitsch, I think it is an excellent point. Our previous risk models weighted morbidity and mortality equally, which I do not think was appropriate. I think we have moved forward, but your points are very well taken and may improve

our ability to predict patient outcomes. Unfortunately, this would require changing the database structure, but it is feasible and worth considering.

DR WILLIAM COOKSEY (Dallas, TX): Why continue with the three-star ratings rather than attempt to move to something a little more granular like a five-star rating that teases out a little better the performance of the facilities?

DR KOZOWER: I think that is a great question, and we have actually been talking about this a lot over the last week. A critical consideration in any ranking is your ability to classify programs correctly. At our hospital, the University of Virginia, we participate with the University HealthSystem Consortium, which ranks hospitals into deciles. As we heard in the discussion of Dr Allen's presentation earlier this morning, one measurement period you can be in the eighth decile and the next you can be in the second decile but there may be no statistical difference between them. So I think that is why we have to be very careful that we are classifying programs correctly, particularly if we are going to start an effort for public reporting and label a program as a

potential one star. We have used 95% Bayesian credible intervals to create our groups, and to call someone a one-star or three-star program, you have to be statistically different from the average composite score. Using these intervals, 90% of programs are two-star or performing at an average/expected level. If you tried to increase the number of groups, we would have even worse discrimination.

DR STEPHEN CASSIVI (Rochester, MN): Benj, congratulations. Your workforce continues to make grade strides. This is about lung cancer, though. I would also note that you are looking into 5-year long-term survival. I think that is going to be the key factor in this composite score. I would encourage you to continue to move toward adding that in, because that is really where the importance of this procedure lies. It is in the long-term outcomes for these cancer patients. Thanks.

DR KOZOWER: Yes, we completely agree. Five-year survival for lung and esophageal cancer resection has been added to the database in 2015 and will be a critical outcome variable for future analyses.