NATIONAL QUALITY FORUM

Measure Submission and Evaluation Worksheet 5.0

This form contains the information submitted by measure developers/stewards, organized according to NQF's measure evaluation criteria and process. The evaluation criteria, evaluation guidance documents, and a blank online submission form are available on the submitting standards web page.

NQF #: 1790 NQF Project: Cancer Project

(for Endorsement Maintenance Review)
Original Endorsement Date: Most Recent Endorsement Date:

BRIEF MEASURE INFORMATION

De.1 Measure Title: Risk-Adjusted Morbidity and Mortality for Lung Resection for Lung Cancer

Co.1.1 Measure Steward: Society of Thoracic Surgeons

De.2 Brief Description of Measure: Percentage of patients = 18 years of age undergoing elective lung resection (Open or VATS wedge resection, segmentectomy, lobectomy, bilobectomy, sleeve lobectomy, pneumonectomy) for lung cancer who developed any of the following postoperative complications: reintubation, need for tracheostomy, initial ventilator support > 48 hours, ARDS, pneumonia, pulmonary embolus, bronchopleural fistula, bleeding requiring reoperation, myocardial infarction or operative mortality.

2a1.1 Numerator Statement: Number of patients = 18 years of age undergoing elective lung resection for lung cancer who developed any of the following postoperative complications: reintubation, need for tracheostomy, initial ventilator support > 48 hours, ARDS, pneumonia, pulmonary embolus, bronchopleural fistula, bleeding requiring reoperation, myocardial infarction or operative mortality.

2a1.4 Denominator Statement: Number of patients = 18 years of age undergoing elective lung resection for lung cancer.

2a1.8 Denominator Exclusions: Emergency procedures

1.1 Measure Type: Outcome
2a1. 25-26 Data Source: Electronic Clinical Data, Electronic Clinical Data : Electronic Health Record, Electronic Clinical Data : Registry, Paper Records
2a1.33 Level of Analysis: Clinician : Group/Practice, Clinician : Team, Facility

1.2-1.4 Is this measure paired with another measure? No

De.3 If included in a composite, please identify the composite measure (*title and NQF number if endorsed*):
n/a

STAFF NOTES (*issues or questions regarding any criteria*)

Comments on Conditions for Consideration:

Is the measure untested? Yes☐ No☐ If untested, explain how it meets criteria for consideration for time-limited endorsement:

1a. Specific national health goal/priority identified by DHHS or NPP addressed by the measure (*check De.5*):
5. Similar/related endorsed or submitted measures (*check 5.1*):
Other Criteria:

Staff Reviewer Name(s):

1. IMPACT, OPPORTUITY, EVIDENCE - IMPORTANCE TO MEASURE AND REPORT

Importance to Measure and Report is a threshold criterion that must be met in order to recommend a measure for endorsement. All

three subcriteria must be met to pass this criterion. See guidance on evidence.
Measures must be judged to be important to measure and report in order to be evaluated against the remaining criteria.
(evaluation criteria)

1a. High Impact: H□ M□ L□ I □
(*The measure directly addresses a specific national health goal/priority identified by DHHS or NPP, or some other high impact aspect of healthcare.*)

De.4 Subject/Topic Areas (*Check all the areas that apply*): Cancer, Cancer : Lung, Esophageal, Surgery, Surgery : Thoracic
De.5 Cross Cutting Areas (*Check all the areas that apply*): Safety, Safety : Complications

1a.1 Demonstrated High Impact Aspect of Healthcare: Affects large numbers, A leading cause of morbidity/mortality, Frequently performed procedure, Patient/societal consequences of poor quality

1a.2 If "Other," please describe:

1a.3 Summary of Evidence of High Impact (*Provide epidemiologic or resource use data*):
Lung cancer is the second most common cancer [1]. An estimated 226,160 new cases of lung cancer are expected in 2012, accounting for about 14% of cancer diagnoses. Lung cancer accounts for more deaths than any other cancer in both men and women[2].

1a.4 Citations for Evidence of High Impact cited in 1a.3: 1. National Cancer Institute website at: http://www.cancer.gov/aboutnci/servingpeople/cancer-statistics/snapshots
2. American Cancer Society: http://www.cancer.org/Research/CancerFactsFigures/CancerFactsFigures/cancer-facts-figures-2012

1b. Opportunity for Improvement: H□ M□ L□ I □
(*There is a demonstrated performance gap - variability or overall less than optimal performance*)

1b.1 Briefly explain the benefits (improvements in quality) envisioned by use of this measure:
Providing outcomes data to participating thoracic surgery sites allows benchmarking of practice group results against the STS national results and allows demonstration of improvement when QI efforts are undertaken. These outcomes data aid clinicians and patients in making informed clinical decisions and also compare risk-adjusted outcomes for quality improvement purposes.

1b.2 Summary of Data Demonstrating Performance Gap (*Variation or overall less than optimal performance across providers*):
[*For Maintenance* – *Descriptive statistics for performance results for this measure - distribution of scores for measured entities by quartile/decile, mean, median, SD, min, max, etc.*]
The endpoint of mortality or major morbidity occurred in 8.6% of eligible patients. Hospital-specific estimates of SIR for mortality or morbidity varied four-fold, with a range of 0.52% to 2.18%.
Dates: January 1, 2008 – December 31, 2010

Data/Sample: The population included 22,677 records from 174 hospitals. Hospital-specific sample sizes ranged from 1 to 883 records per hospital.

Distribution of hospital-specific estimates of standardized incidence ratio (SIR) for composite of mortality and morbidity:

Minimum	0.52
1st quartile	0.86
Median	1.04
Mean	1.06
3rd quartile	1.20
Maximum	2.18

1b.3 Citations for Data on Performance Gap: [*For Maintenance* – *Description of the data or sample for measure results reported in 1b.2 including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included*]
Increased health care utilization may be a useful inverse surrogate for surgical quality, but little is known about health care utilization after lung resection. Risk-adjustment may be necessary if risk factors for utilization vary across providers, and,

importantly, risk-adjustment models may also have to account for nonclinical determinants of increased utilization. The descriptive information from this study provides a framework for future investigations and discussions about how best to use measures of increased health care utilization for surgical quality improvement [1].
1. Farjah F et al. Health Care utilization among surgically treated Medicare beneficiaries with lung cancer. Ann Thorac Surg 2009;88:1749-56.

1b.4 Summary of Data on Disparities by Population Group: [*For Maintenance –Descriptive statistics for performance results for this measure by population group*]
The Disparities in Care Table in the attachment summarizes proportions of patients with mortality or morbidity by race. Even though estimates of these proportions differ somewhat by race, the 95% confidence intervals overlap.

1b.5 Citations for Data on Disparities Cited in 1b.4: [*For Maintenance – Description of the data or sample for measure results reported in 1b.4 including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included*]
LaPar DJ et al. Gender, race and socioeconomic status affects outcome after lung cancer resections in the United States. Ann Thorac Surg 2011;92:434-9.

1c. Evidence (*Measure focus is a health outcome OR meets the criteria for quantity, quality, consistency of the body of evidence.*)
Is the measure focus a health outcome? Yes□ No□ If not a health outcome, rate the body of evidence.

Quantity: H□ M□ L□ I □ Quality: H□ M□ L□ I □ Consistency: H□ M□ L□ I □

Quantity	Quality	Consistency	Does the measure pass subcriterion1c?
M-H	M-H	M-H	Yes□
L	M-H	M	Yes□ IF additional research unlikely to change conclusion that benefits to patients outweigh harms: otherwise No□
M-H	L	M-H	Yes□ IF potential benefits to patients clearly outweigh potential harms: otherwise No□
L-M-H	L-M-H	L	No □

Health outcome – rationale supports relationship to at least one healthcare structure, process, intervention, or service	Does the measure pass subcriterion1c? Yes□ IF rationale supports relationship

1c.1 Structure-Process-Outcome Relationship *(Briefly state the measure focus, e.g., health outcome, intermediate clinical outcome, process, structure; then identify the appropriate links, e.g., structure-process-health outcome; process- health outcome; intermediate clinical outcome-health outcome)*:
The measure focus is health outcome. Reduced morbidity and mortality following postoperative complications is a goal when elective lung resections are performed.
With all hospitals included, the Bayesian estimate of the reliability measure (squared correlation between a measurement and the true value) is 0.51 with the 95% Bayesian probability interval (0.40, 0.61). Since the lower limit of this interval is 0.40, we may be highly confident (probability = 97.5%) that the true reliability is at least 0.40.), this reliability measure it is 0.55 (0.43, 0.61) for 160 hospitals performing at least 10 procedures, and it is 0.77 (0.62,0.88) for 35 hospitals with 200 or more procedures performed.
When estimated with 3 years of data, the proposed lung cancer morbidity and mortality measure is reliable enough to be useful in the context of feedback reporting for internal quality improvement initiatives.

1c.2-3 Type of Evidence *(Check all that apply)*:
Other, Selected individual studies (rather than entire body of evidence)
..

1c.4 Directness of Evidence to the Specified Measure *(State the central topic, population, and outcomes addressed in the body of evidence and identify any differences from the measure focus and measure target population)*:
Measure includes only elective lung resections, for lung cancer patients.

1c.5 Quantity of Studies in the Body of Evidence *(Total number of studies, not articles)*: The body of evidence includes 84 studies.

1c.6 Quality of Body of Evidence (*Summarize the certainty or confidence in the estimates of benefits and harms to patients across studies in the body of evidence resulting from study factors. Please address: a) study design/flaws; b) directness/indirectness of the evidence to this measure (e.g., interventions, comparisons, outcomes assessed, population included in the evidence); and c) imprecision/wide confidence intervals due to few patients or events*): a) Presented data are from a high quality database with substantial number of patients (n=22,677) within 36 months' time period.
b) Population and outcome measure are of direct relevance.
c) Bayesian 95% probability interval for reliability measure (number between 0 and 1) is provided and it is satisfactorily narrow with length of approximately 0.2.

1c.7 Consistency of Results across Studies (*Summarize the consistency of the magnitude and direction of the effect*): There is an approximate 4 fold variation in outcome measure following lung cancer resections across hospitals with standardized incidence ration (SIR) ranging between 0.52 and 2.18.

1c.8 Net Benefit (*Provide estimates of effect for benefit/outcome; identify harms addressed and estimates of effect; and net benefit - benefit over harms*):
Reduction in deaths and/or complications after lung cancer resection will be a benefit.

1c.9 Grading of Strength/Quality of the Body of Evidence. Has the body of evidence been graded? No

1c.10 If body of evidence graded, identify the entity that graded the evidence including balance of representation and any disclosures regarding bias: n/a

1c.11 System Used for Grading the Body of Evidence: Other

1c.12 If other, identify and describe the grading scale with definitions: n/a

1c.13 Grade Assigned to the Body of Evidence: n/a

1c.14 Summary of Controversy/Contradictory Evidence: n/a

1c.15 Citations for Evidence other than Guidelines(*Guidelines addressed below*):
n/a

1c.16 Quote verbatim, <u>the specific guideline recommendation</u> (*Including guideline # and/or page #*):
n/a

1c.17 Clinical Practice Guideline Citation: n/a

1c.18 National Guideline Clearinghouse or other URL: n/a

1c.19 Grading of Strength of Guideline Recommendation. Has the recommendation been graded? No

1c.20 If guideline recommendation graded, identify the entity that graded the evidence including balance of representation and any disclosures regarding bias:

1c.21 System Used for Grading the Strength of Guideline Recommendation: Other

1c.22 If other, identify and describe the grading scale with definitions: n/a

1c.23 Grade Assigned to the Recommendation: n/a

1c.24 Rationale for Using this Guideline Over Others: n/a

Based on the NQF descriptions for rating the evidence, what was the <u>developer's assessment</u> of the quantity, quality, and

consistency of the body of evidence?
1c.25 Quantity: High 1c.26 Quality: High1c.27 Consistency: High

Was the threshold criterion, *Importance to Measure and Report*, met?
(*1a & 1b must be rated moderate or high and 1c yes*) Yes□ No□
Provide rationale based on specific subcriteria:

For a new measure if the Committee votes NO, then STOP.
For a measure undergoing endorsement maintenance, if the Committee votes NO because of 1b. (no opportunity for improvement), it may be considered for continued endorsement and all criteria need to be evaluated.

2. RELIABILITY & VALIDITY - SCIENTIFIC ACCEPTABILITY OF MEASURE PROPERTIES

Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. (evaluation criteria)
Measure testing must demonstrate adequate reliability and validity in order to be recommended for endorsement. Testing may be conducted for data elements and/or the computed measure score. Testing information and results should be entered in the appropriate field. Supplemental materials may be referenced or attached in item 2.1. See guidance on measure testing.

S.1 Measure Web Page *(In the future, NQF will require measure stewards to provide a URL link to a web page where current detailed specifications can be obtained*). Do you have a web page where current detailed specifications for this measure can be obtained? Yes

S.2 If yes, provide web page URL: http://www.sts.org/sites/default/files/documents/STSThoracicDataSpecsV2_2.pdf

2a. RELIABILITY. Precise Specifications and Reliability Testing: H□ M□ L□ I □

2a1. Precise Measure Specifications. (*The measure specifications precise and unambiguous.*)

2a1.1 Numerator Statement *(Brief, narrative description of the measure focus or what is being measured about the target population, e.g., cases from the target population with the target process, condition, event, or outcome)*:
Number of patients = 18 years of age undergoing elective lung resection for lung cancer who developed any of the following postoperative complications: reintubation, need for tracheostomy, initial ventilator support > 48 hours, ARDS, pneumonia, pulmonary embolus, bronchopleural fistula, bleeding requiring reoperation, myocardial infarction or operative mortality.

2a1.2 Numerator Time Window *(The time period in which the target process, condition, event, or outcome is eligible for inclusion)*:
During hospitalization regardless of length of stay or within 30 days of surgery if discharged from the hospital.

2a1.3 Numerator Details *(All information required to identify and calculate the cases from the target population with the target process, condition, event, or outcome such as definitions, codes with descriptors, and/or specific data collection items/responses*:
Number of patients undergoing elective lung resection for lung cancer for whom:

1. Postoperative events (POEvents - STS GTS Database, v 2.2, sequence number 1710) is marked “Yes” and one of the following items is marked:
a. Reintubation (Reintube - STS GTS Database, v 2.2, sequence number 1850)
b. Need for tracheostomy (Trach - STS GTS Database, v 2.2, sequence number 1860)
c. Initial ventilator support > 48 hours (Vent- STS GTS Database, v 2.2, sequence number 1840)
d. Adult Respiratory Distress Syndrome (ARDS - STS GTS Database, v 2.2, sequence number 1790)
e. Pneumonia (Pneumonia - STS GTS Database, v 2.2, sequence number 1780)
f. Pulmonary Embolus (PE - STS GTS Database, v 2.2, sequence number 1820)
g. Bronchopleural Fistula (Bronchopleural - STS GTS Database, v 2.2, sequence number 1810)
h. Myocardial infarction (MI - STS GTS Database, v 2.2, sequence number 1900)

Or

2. Unexpected return to the operating room (ReturnOR - STS GTS Database, Version 2.2, sequence number 1720) is marked “yes” and primary reason for return to OR (ReturnORRsn – STS GTS Database, Version 2.2, sequence number 1730) is

marked "bleeding"

Or

3. One of the following fields is marked "dead"
a. Discharge status (MtDCStat - STS GTS Database, Version 2.2, sequence number 2200);
b. Status at 30 days after surgery (Mt30Stat - STS GTS Database, Version 2.2, sequence number 2240)

Please see STS General Thoracic Surgery Database Data Collection Form, Version 2.2- http://www.sts.org/sites/default/files/documents/STSThoracicDCF_V2_2_MajorProc_Annotated_0.pdf

2a1.4 Denominator Statement *(Brief, narrative description of the target population being measured)*:
Number of patients = 18 years of age undergoing elective lung resection for lung cancer.

2a1.5 Target Population Category *(Check all the populations for which the measure is specified and tested if any)*: Adult/Elderly Care

2a1.6 Denominator Time Window *(The time period in which cases are eligible for inclusion)*:
36 months

2a1.7 Denominator Details (*All information required to identify and calculate the target population/denominator such as definitions, codes with descriptors, and/or specific data collection items/responses)*:
1. Lung cancer (LungCancer - STS GTS Database, v 2.2, sequence number 830) is marked "yes" and Category of Disease – Primary (CategoryPrim - STS GTS Database, v 2.2, sequence number 1300) is marked as one of the following:

(ICD-9, ICD-10)
Lung cancer, main bronchus, carina (162.2, C34.00)
Lung cancer, upper lobe (162.3, C34.10)
Lung cancer, middle lobe (162.4, C34.2)
Lung cancer, lower lobe (162.5, C34.30)
Lung cancer, location unspecified (162.9, C34.90)

2. Patient has lung cancer (as defined in #1 above) and primary procedure is one of the following CPT codes:

Thoracoscopy, surgical; with lobectomy (32663)
Thoracoscopy with therapeutic wedge resection (eg mass or nodule) initial, unilateral (3266X)
Thoracoscopy with therapeutic wedge resection (eg mass or nodule) each additional resection, ipsilateral (3266X1)
Thoracoscopy with diagnostic wedge resection followed by anatomic lung resection (3266X2)
Thoracoscopy with removal of a single lung segment (segmentectomy) (3266X4)
Thoracoscopy with removal of two lobes (bilobectomy) (3266X3)
Thoracoscopy with removal of lung, pneumonectomy (3266X5)
Thoracotomy with therapeutic wedge resection (eg mass nodule) initial (3250X)
Thoracotomy with therapeutic wedge resection (eg mass nodule) each additional resection, ipsilateral (+3250X1)
Thoracotomy with diagnostic wedge resection followed by anatomic lung resection (+3250X2)
Removal of lung, total pneumonectomy; (32440)
Removal of lung, sleeve (carinal) pneumonectomy (32442)
Removal of lung, total pneumonectomy; extrapleural (32445)
Removal of lung, single lobe (lobectomy) (32480)
Removal of lung, two lobes (bilobectomy) (32482)
Removal of lung, single segment (segmentectomy) (32484)
Removal of lung, sleeve lobectomy (32486)
Removal of lung, completion pneumonectomy (32488)
Resection of apical lung tumor (e.g., Pancoast tumor), including chest wall resection, without chest wall reconstruction(s) (32503)
Resection of apical lung tumor (e.g., Pancoast tumor), including chest wall resection, with chest wall reconstruction (32504)

3. Status of Operation (Status - STS General Thoracic Surgery Database, Version 2.2, sequence number 1420) is marked as "Elective"

4. Only analyze the first operation of the hospitalization meeting criteria 1-3

2a1.8 Denominator Exclusions *(Brief narrative description of exclusions from the target population)*:
Emergency procedures

2a1.9 Denominator Exclusion Details *(All information required to identify and calculate exclusions from the denominator such as definitions, codes with descriptors, and/or specific data collection items/responses)*:
n/a

2a1.10 Stratification Details/Variables *(All information required to stratify the measure results including the stratification variables, codes with descriptors, definitions, and/or specific data collection items/responses)*:
n/a

2a1.11 Risk Adjustment Type *(Select type. Provide specifications for risk stratification in 2a1.10 and for statistical model in 2a1.13)*: Statistical risk model 2a1.12 If "Other," please describe:

2a1.13 Statistical Risk Model and Variables *(Name the statistical method - e.g., logistic regression and list all the risk factor variables. Note - risk model development should be addressed in 2b4.)*:
Bayesian hierarchical modeling was used to assess the statistical reliability of hospital-specific standardized incidence ratio (SIR) estimates derived from the January 1, 2008 – December 31, 2010 STS data. All hospitals regardless of sample size were included in the estimation of model parameters. Reliability measures were initially calculated including all the hospitals and were subsequently calculated in subsets of hospitals having at least 10, 20, 30, 50, 100, or 200 eligible cases.
Three separate multivariable risk models were constructed (mortality, major morbidity, and composite mortality or major morbidity). The risk-adjustment models created for this measure and study have excellent performance characteristics and identify important predictors of mortality and major morbidity for lung cancer resections. These models may be used to inform clinical decisions and to compare risk-adjusted outcomes for quality improvement purposes. For additional information see the attachment:

Kozower BD, Sheng S, O'Brien SM, Liptay MJ, Lau CL, Jones DR, Shahian DM, Wright CD. STS Database Risk Models: Predictors of Mortality and Major Morbidity for Lung Cancer Resection. Ann Thorac Surg. 2010;90:875–83.

2a1.14-16 Detailed Risk Model Available at Web page URL (or attachment). Include coefficients, equations, codes with descriptors, definitions, and/or specific data collection items/responses. Attach documents only if they are not available on a webpage and keep attached file to 5 MB or less. NQF strongly prefers you make documents available at a Web page URL. Please supply login/password if needed:
Attachment
Kozower et al.pdf

2a1.17-18. Type of Score: Rate/proportion

2a1.19 Interpretation of Score *(Classifies interpretation of score according to whether better quality is associated with a higher score, a lower score, a score falling within a defined interval, or a passing score)*: Better quality = Lower score

2a1.20 Calculation Algorithm/Measure Logic*(Describe the calculation of the measure score as an ordered sequence of steps including identifying the target population; exclusions; cases meeting the target process, condition, event, or outcome; aggregating data; risk adjustment; etc.)*:
Target population is patients 18 years of age or older undergoing elective lung resection for lung cancer. Emergency procedures were excluded. Outcome is occurrence of postoperative complications: reintubation, need for tracheostomy, initial ventilator support > 48 hours, ARDS, pneumonia, pulmonary embolus, bronchopleural fistula, bleeding requiring reoperation, myocardial infarction or operative mortality. Analysis considered 22,677 patients with procedures between 01/01/2008 and 12/31/2010 (36

months). Risk adjustment was achieved with a Bayesian hierarchical model with composite of the above postoperative complications as the outcome. The measure score was estimated with this model. For additional information review risk model in attachment.

2a1.21-23 Calculation Algorithm/Measure Logic Diagram URL or attachment:

2a1.24 Sampling (Survey) Methodology. If measure is based on a sample (or survey), provide instructions for obtaining the sample, conducting the survey and guidance on minimum sample size (response rate):
n/a

2a1.25 Data Source *(Check all the sources for which the measure is specified and tested*). If other, please describe:
Electronic Clinical Data, Electronic Clinical Data : Electronic Health Record, Electronic Clinical Data : Registry, Paper Records

2a1.26 Data Source/Data Collection Instrument (*Identify the specific data source/data collection instrument, e.g. name of database, clinical registry, collection instrument, etc.*): STS General Thoracic Surgery Database, Version 2.2

2a1.27-29 Data Source/data Collection Instrument Reference Web Page URL or Attachment: URL
Data Collection Form-http://www.sts.org/sites/default/files/documents/STSThoracicDCF_V2_2_MajorProc_Annotated_0.pdf

2a1.30-32 Data Dictionary/Code Table Web Page URL or Attachment:
URL
http://www.sts.org/sites/default/files/documents/STSThoracicDataSpecsV2_2.pdf

2a1.33 Level of Analysis (*Check the levels of analysis for which the measure is specified and tested*): Clinician : Group/Practice, Clinician : Team, Facility

2a1.34-35 Care Setting *(Check all the settings for which the measure is specified and tested)*: Hospital/Acute Care Facility

2a2. Reliability Testing. (*Reliability testing was conducted with appropriate method, scope, and adequate demonstration of reliability*.)

2a2.1 Data/Sample *(Description of the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)*:
The analysis population consisted of all STS records for patients meeting measure inclusion criteria who had their surgery during January 1, 2008 – December 31, 2010. The population included 22,677 records from 174 hospitals. Hospital-specific sample sizes ranged from 1 to 883 records per hospital.

2a2.2 Analytic Method *(Describe method of reliability testing & rationale)*:
All hospitals regardless of sample size were included in the estimation of Bayesian model parameters. Reliability measures were initially calculated including all the hospitals and were subsequently calculated in subsets of hospitals with specified minimum number of performed procedures.
Reliability is conventionally defined as the proportion of variation in a performance measure that is due to true between-hospital differences (i.e., signal) as opposed to random statistical fluctuations (i.e., noise). A mathematically equivalent definition is the squared correlation between a measurement and the true value. We used two different approaches to operationalize the above definitions of reliability in a Bayesian statistical framework. First, the proportion of variance in the Bayesian posterior distribution attributed to differences between hospitals was calculated. An alternative closely related reliability measure was computed; i.e. the squared correlation between each hospital's estimated performance measure (the estimated SIR) and the true value (estimated using Bayesian inference methods).

2a2.3 Testing Results *(Reliability statistics, assessment of adequacy in the context of norms for the test conducted)*:
Technical Details

Prior to estimating reliability, the numerical value of SIR was estimated for each hospital under the model described by Kozower et al. (2010). The estimated proportion of variance in the Bayesian posterior distribution attributed to differences between hospitals was calculated and this was the first reliability measure (higher proportion implies higher reliability). This reliability measure is 0.50 when all 174 hospitals were considered (including those with very few procedures), it is 0.54 for 160 hospitals with at least 10 procedures, and it is 0.75 for 35 hospitals with 200 or more procedures performed.
The second closely related reliability measure was defined as the estimated squared correlation between the set of hospital-specific estimates of SIR and the corresponding unknown true values (estimated using Bayesian inference methods). A 95% Bayesian probability interval for this reliability measure was obtained. With all 174 hospitals included, the estimate of the second reliability measure is 0.51 and the 95% Bayesian probability interval (0.40, 0.61), it is 0.55 (0.43, 0.61) for 160 hospitals performing at least 10 procedures, and it is 0.77 (0.62,0.88) for 35 hospitals with 200 or more procedures performed.
In summary, when estimated with 3 years of data, the proposed lung cancer morbidity and mortality measure is reliable enough to be useful in the context of feedback reporting for internal quality improvement initiatives. Public reporting may be considered for a subset of hospitals with a larger case volume.

2b. VALIDITY. Validity, Testing, including all Threats to Validity: H□ M□ L□ I □

2b1.1 Describe how the measure specifications *(measure focus, target population, and exclusions)* are consistent with the evidence cited in support of the measure focus (*criterion 1c)* and identify any differences from the evidence:
The specifications are quite standard-operable lung cancer patients undergoing elective resection.

2b2. Validity Testing. (*Validity testing was conducted with appropriate method, scope, and adequate demonstration of validity.)*

2b2.1 Data/Sample *(Description of the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)*:
STS General Thoracic Surgery Database. The analysis population consisted of all STS records for patients meeting measure inclusion criteria who had their surgery during January 1, 2008 – December 31, 2010. The population included 22,677 records from 174 hospitals. Hospital-specific sample sizes ranged from 1 to 883 records per hospital.

2b2.2 Analytic Method *(Describe method of validity testing and rationale; if face validity, describe systematic assessment)*:
When data arrive at the data warehouse, they are checked carefully for logical inconsistencies, missing required fields, and parent/child variable relationship violations. Any inconsistencies or violations are communicated to participants in the detailed Data Quality Report that is generated automatically following each harvest file submission. Upon receipt of the Data Quality Report, participants are given an opportunity to correct the data, which substantially improves the quality and completeness of the data submitted for analysis. If the data inconsistencies are not changed by the participant prior to harvest close, the data warehouse performs consistency edits and/or parent/child edits on the data in order for them to be analyzable. Participants are informed of such edits to their data in the Data Quality Report.

Since 2010, Telligan (formerly IFMC) has conducted audits of the STS General Thoracic Surgery Database on the Society´s behalf to evaluate the accuracy, consistency and comprehensiveness of data collection, which has validated the integrity of the data. Auditors havevalidated case inclusion and twenty lobectomy cases were randomly chosen for review of thirty-three individual data elements. The auditors have abstracted each designated medical record to validate data elements previously submitted to the STS data warehouse. Agreement rates have been calculated for each of the 33 elements as well as for an overall agreement rate. Five sites were randomly selected for the first audit, which took place in the fall 2010. In 2011, 10 sites were audited.

In addition, validity was confirmed and is regularly assessed by an expert panel of thoracic surgeons assembled by the STS General Thoracic Surgery Database Task Force, the STS Task Force on Quality Initiatives, and the STS Workforce on National Databases.

2b2.3 Testing Results *(Statistical results, assessment of adequacy in the context of norms for the test conducted; if face validity, describe results of systematic assessment)*:
STS audited 5% of participants in the General Thoracic Surgery Database in 2011 using an independent auditing firm. The sites were randomly selected and audited for data completeness and accuracy. Auditors compared case logs at each facility and cases submitted to the STS GTSD to assess completeness of data submission. There was consistent agreement across all participants for data completeness. Data accuracy was assessed by reabstraction of 20 randomly chosen lobectomy cases, comparing 33 data elements in the medical chart with the data file submitted to the STS GTSD. The agreement rate was 94.61% for overall data accuracy, with a range in agreement from 76.5% to 95.5% in 2010. This same range in agreement was 88.8% to 97.5% in 2011.

Theoverall agreement across all 33 elements was 89.9% in2010and 94.6% in 2011.

POTENTIAL THREATS TO VALIDITY. (*All potential threats to validity were appropriately tested with adequate results.*)

2b3. Measure Exclusions. (*Exclusions were supported by the clinical evidence in 1c or appropriately tested with results demonstrating the need to specify them.*)

2b3.1 Data/Sample for analysis of exclusions *(Description of the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)*:
n/a

2b3.2 Analytic Method (*Describe type of analysis and rationale for examining exclusions, including exclusion related to patient preference)*:
n/a

2b3.3 Results *(Provide statistical results for analysis of exclusions, e.g., frequency, variability, sensitivity analyses):*
n/a

2b4. Risk Adjustment Strategy. (*For outcome measures, adjustment for differences in case mix (severity) across measured entities was appropriately tested with adequate results.*)

2b4.1 Data/Sample *(Description of the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)*:
The STS General Thoracic Surgery Database was queried for all patients treated with resection for primary lung cancer between January 1, 2002 and June 30, 2008 (i.e., 18,800 lung cancer resections at 111 participating sites). Three separate multivariable risk models were constructed (mortality, major morbidity, and composite mortality or major morbidity). [1]

1. Kozower BD, Sheng S, O'Brien SM, Liptay MJ, Lau CL, Jones DR, Shahian DM, Wright CD. STS Database Risk Models: Predictors of Mortality and Major Morbidity for Lung Cancer Resection. Ann Thorac Surg. 2010;90:875–83.

2b4.2 Analytic Method (*Describe methods and rationale for development and testing of risk model or risk stratification including selection of factors/variables)*:
Bayesian hierarchical modeling was used to estimate hospital-specific standardized incidence ratio (SIR) and a 95% Bayesian probability interval for SIR for each of 174 hospitals. This analytic method is the same method used in Kozower, etc.

1. Kozower BD, Sheng S, O'Brien SM, Liptay MJ, Lau CL, Jones DR, Shahian DM, Wright CD. STS Database Risk Models: Predictors of Mortality and Major Morbidity for Lung Cancer Resection. Ann Thorac Surg. 2010;90:875–83.

2b4.3 Testing Results *(Statistical risk model: Provide quantitative assessment of relative contribution of model risk factors; risk model performance metrics including cross-validation discrimination and calibration statistics, calibration curve and risk decile plot, and assessment of adequacy in the context of norms for risk models. Risk stratification: Provide quantitative assessment of relationship of risk factors to the outcome and differences in outcomes among the strata)*:
Discrimination of the model was assessed by the area under the receiver operating characteristic curve, also known as the C-statistic, from each of the ten imputation datasets. The statistical significance of the difference in the observed to expect numbers of the outcome was assessed using the Hosmer-Lemeshow goodness-of-fit test. The bootstrap-adjusted C-statistics is 0.69. The Hosmer-Lemeshow goodness-of-fit p-values=0.47 demonstrates that the model estimates fit the data at an acceptable level. Kozower BD, Sheng S, O'Brien SM, Liptay MJ, Lau CL, Jones DR, Shahian DM, Wright CD. STS Database Risk Models: Predictors of Mortality and Major Morbidity for Lung Cancer Resection. Ann Thorac Surg. 2010;90:875–83.

2b4.4 If outcome or resource use measure is not risk adjusted, provide rationale and analyses to justify lack of adjustment: n/a

2b5. Identification of Meaningful Differences in Performance. (*The performance measure scores were appropriately analyzed and discriminated meaningful differences in quality.*)

2b5.1 Data/Sample *(Describe the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)*:

The analysis population consisted of all STS records for patients meeting measure inclusion criteria who had their surgery during January 1, 2008 – December 31, 2010. The population included 22,677 records from 174 hospitals. Hospital-specific sample sizes ranged from 1 to 883 records per hospital with mean 175 and median 130 records per hospital, and interquartile range (36, 175).

2b5.2 Analytic Method *(Describe methods and rationale to identify statistically significant and practically/meaningfully differences in performance)*:
Bayesian hierarchical modeling was used to estimate hospital-specific standardized incidence ratio (SIR) and a 95% Bayesian probability interval for SIR for each of 174 hospitals.

2b5.3 Results *(Provide measure performance results/scores, e.g., distribution by quartile, mean, median, SD, etc.; identification of statistically significant and meaningfully differences in performance)*:
The Plot under the Results section of the attachment displays estimated SIR and corresponding 95% Bayesian probability interval for each of 174 hospitals. Hospitals are ordered according to the increasing SIR estimate. There are meaningful differences between the best performing and the worst performing hospitals.

2b6. Comparability of Multiple Data Sources/Methods. (*If specified for more than one data source, the various approaches result in comparable scores.*)

2b6.1 Data/Sample *(Describe the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)*:
n/a

2b6.2 Analytic Method *(Describe methods and rationale for testing comparability of scores produced by the different data sources specified in the measure)*:
n/a

2b6.3 Testing Results *(Provide statistical results, e.g., correlation statistics, comparison of rankings; assessment of adequacy in the context of norms for the test conducted)*:
n/a

2c. Disparities in Care: H□ M□ L□ I □ NA□ (*If applicable, the measure specifications allow identification of disparities.*)

2c.1 If measure is stratified for disparities, provide stratified results *(Scores by stratified categories/cohorts)*: Under the Disparities in Care section of the attachment, the table summarizes proportions of patients with mortality or morbidity by race. Even though estimates of these proportions differ somewhat by race, the 95% confidence intervals overlap.

2c.2 If disparities have been reported/identified (e.g., in 1b), but measure is not specified to detect disparities, please explain:
n/a

2.1-2.3 Supplemental Testing Methodology Information:
Attachment
sections 1b.2, 1b.4, 2a2.2, 2a2.3, 2b5.3, 2c.1.doc

Steering Committee: Overall, was the criterion, *Scientific Acceptability of Measure Properties*, met?
(*Reliability and Validity must be rated moderate or high*) Yes□ No□
Provide rationale based on specific subcriteria:

If the Committee votes No, STOP

3. USABILITY

Extent to which intended audiences (e.g., consumers, purchasers, providers, policy makers) can understand the results of the measure and are likely to find them useful for decision making. (evaluation criteria)

C.1 Intended Purpose/ Use *(Check all the purposes and/or uses for which the measure is intended)*:   Public Reporting, Quality

Improvement (Internal to the specific organization), Quality Improvement with Benchmarking (external benchmarking to multiple organizations)

3.1 Current Use *(Check all that apply; for any that are checked, provide the specific program information in the following questions)*: Quality Improvement with Benchmarking (external benchmarking to multiple organizations), Quality Improvement (Internal to the specific organization)

3a. Usefulness for Public Reporting: H□ M□ L□ I □
(*The measure is meaningful, understandable and useful for public reporting.*)

3a.1. Use in Public Reporting - disclosure of performance results to the public at large *(If used in a public reporting program, provide name of program(s), locations, Web page URL(s)).* If not publicly reported in a national or community program, state the reason AND plans to achieve public reporting, potential reporting programs or commitments, and timeline, e.g., within 3 years of endorsement: [*For Maintenance* – *If not publicly reported, describe progress made toward achieving disclosure of performance results to the public at large and expected date for public reporting; provide rationale why continued endorsement should be considered.*]
STS is diligently working on operationalizing the public reporting of the STS adult cardiac surgery measures. In our efforts to operationalize public reporting, the STS has a Quality Initiatives Task Force and a Task Force on Quality Measurement that are working to develop a public report card that will be consumer centric. There are now more than 400 Adult Cardiac Surgery Database (ACSD) participants who have voluntarily consented to be a part of the STS Public Reporting Online and 366 for the Consumer Reports public reporting initiative of STS data. Public reporting remains a top priority for the Society, and STS is striving for even stronger involvement among Database participants. Discussions have begun regarding STS's timeline for its general thoracic surgery measures. STS intends to publicly report this measure within 2-3 years. Recently, STS upgraded the GTSD data specifications that now include several important process measures that will be useful as building blocks for a composite measure report on the General Thoracic Surgery Database, that we envision reporting to the public. This pulmonary resection morbidity and mortality measure will be one very important component of the overall composite measure.

3a.2.Provide a rationale for why the measure performance results are meaningful, understandable, and useful for public reporting. If usefulness was demonstrated (e.g., focus group, cognitive testing), describe the data, method, and results: STS's combined mortality and morbidity model for pulmonary resection for lung cancer is important and appropriate for the following reasons: 1.) lung cancer resection is the most common major procedure that a thoracic surgeon performs, 2.) this one procedure provides the most compelling reason to model, risk stratify and allow results comparisons so that underperforming centers identify the rationale to enhance their quality improvement (QI) efforts, 3.) major morbidity is relatively common after lung resection, but mortality, while rare, should be captured as well in an effort to identify adverse events after lung resection, by combining these two outcome measures for reporting, 4.) this measure will be reported in an easy to understand format which plts the participants O/E ratio with the STS median, 25th and 75th percentiles along with 95% confidence intervals. Surgeons easily grasp this result and the visual display powerfully shows them just where they perform compared to their peers on a bi-annual basis. In addition, these risk-adjusted results allow surgeons to benchmark their program and initiative QI efforts, as needed. In providing transparency in our public reporting efforts with this measure, it will help surgeons better compare their patients' outcomes with national benchmarks and we will have better informed consumers of health care.

3.2 Use for other Accountability Functions (payment, certification, accreditation). If used in a public accountability program, provide name of program(s), locations, Web page URL(s): n/a

3b. Usefulness for Quality Improvement: H□ M□ L□ I □
(*The measure is meaningful, understandable and useful for quality improvement.*)

3b.1. Use in QI. If used in quality improvement program, provide name of program(s), locations, Web page URL(s):
[*For Maintenance* – *If not used for QI, indicate the reasons and describe progress toward using performance results for improvement*].
The STS GTSD was established as a voluntary initiative to support the continuous quality improvement efforts of surgeons and hospitals. Participating institutions receive twice-yearly feedback reports that describe each site's results in relation to other database participants. Each site uses these feedback reports to enhance its quality improvement efforts.

3b.2. Provide rationale for why the measure performance results are meaningful, understandable, and useful for quality improvement. If usefulness was demonstrated *(e.g., QI initiative),* describe the data, method and results:

STS is diligently working on operationalizing the public reporting of the STS adult cardiac surgery measures. In our efforts to operationalize the public reporting the STS has a Quality Initiative Task Force and Task Force on Quality Measurement that are working to develop a public report card that will be consumer centric. There are now more than 400 Adult Cardiac Surgery Database (ACSD) participants who have voluntarily consented to be a part of the STS Public Reporting Online and 366 for the Consumer Reports public reporting initiative. Public reporting remains a top priority for the Society, and STS is striving for even stronger involvement among Database participants. Discussions have begun regarding STS's timeline for its general thoracic surgery measures. STS intends to publicly report this measure within 2-3 years.

Overall, to what extent was the criterion, *Usability*, met? H□ M□ L□ I □
Provide rationale based on specific subcriteria:

4. FEASIBILITY

Extent to which the required data are readily available, retrievable without undue burden, and can be implemented for performance measurement. (evaluation criteria)

4a. Data Generated as a Byproduct of Care Processes: H□ M□ L□ I □

4a.1-2 How are the data elements needed to compute measure scores generated? *(Check all that apply).*
Data used in the measure are:
generated by and used by healthcare personnel during the provision of care, e.g., blood pressure, lab value, medical condition, Coded by someone other than person obtaining original information (e.g., DRG, ICD-9 codes on claims), Abstracted from a record by someone other than person obtaining original information (e.g., chart abstraction for quality measure or registry)

4b. Electronic Sources: H□ M□ L□ I □

4b.1 Are the data elements needed for the measure as specified available electronically *(Elements that are needed to compute measure scores are in defined, computer-readable fields)*: ALL data elements are in a combination of electronic sources

4b.2 If ALL data elements are not from electronic sources, specify a credible, near-term path to electronic capture, OR provide a rationale for using other than electronic sources:

4c. Susceptibility to Inaccuracies, Errors, or Unintended Consequences: H□ M□ L□ I □

4c.1 Identify susceptibility to inaccuracies, errors, or unintended consequences of the measurement identified during testing and/or operational use and strategies to prevent, minimize, or detect. If audited, provide results:
This measure may be susceptible to human error (i.e., recording the data elements in the measure inaccurately or not at all).

Both STS and the Duke Clinical Research Institute have a list of database participants making participation in the STS General Thoracic Surgery Database easy to track.

Each participant is responsible for the quality and accuracy of the data they submit to the database. Each participant agrees to the following quality control measures in the participation agreement:

i)"Participant hereby warrants that all data submitted for inclusion in the GTS Database will be accurate and complete, and acknowledges that such data may be subject to independent audit. Participant will use its best efforts to address any data or related deficiencies identified by the independent data warehouse service provider, and agrees to cooperate with and assist STS and its designees in connection with the performance of any independent audit.

ii) Participant warrants that it will take all reasonable steps to avoid the submission of duplicative data for inclusion in the GTS Database, including but not limited to apprising the Director of the STS National Database and the independent data warehouse service provider about any other Participation Agreements in which an individual cardiothoracic surgeon named above or on Schedule A attached hereto (as amended from time to time) is also named.

In addition, the data warehouse and analysis center at Duke Clinical Research Institute, performs a series of internal quality controls on the submitted data.

4d. Data Collection Strategy/Implementation: H□ M□ L□ I □

A.2 Please check if either of the following apply (*regarding proprietary measures*):

4d.1 Describe what you have learned/modified as a result of testing and/or operational use of the measure regarding data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, time and cost of data collection, other feasibility/implementation issues *(e.g., fees for use of proprietary measures)*:

Missing data are sought by the DCRI from participants when the data are initially sent to DCRI for analysis.

Data are collected continuously by the participating sites and harvested by the DCRI twice yearly. Reports are then sent back to the sites about 3 months after a harvest.

No individual patient identifiers are collected by the DCRI.

Data Collection:
Participants of the STS General Thoracic Surgery Database generally have data managers on staff to collect these data. Costs to develop the measure included volunteer thoracic surgeons' time, STS staff time, and DCRI statistician and project management time.

Other fees:
STS General Thoracic Surgery Database participant surgeons pay an annual participant fee of $400 or $500, depending on whether the participant is an STS member or not. As a benefit of STS membership, STS members receive a 25% discount on the fee.

Overall, to what extent was the criterion, *Feasibility*, met? H□ M□ L□ I □
Provide rationale based on specific subcriteria:

OVERALL SUITABILITY FOR ENDORSEMENT

Does the measure meet all the NQF criteria for endorsement? Yes□ No□
Rationale:

If the Committee votes No, STOP.
If the Committee votes Yes, the final recommendation is contingent on comparison to related and competing measures.

5. COMPARISON TO RELATED AND COMPETING MEASURES

If a measure meets the above criteria and there are endorsed or new related measures (either the same measure focus or the same target population) or competing measures (both the same measure focus and the same target population), the measures are compared to address harmonization and/or selection of the best measure before a final recommendation is made.

5.1 If there are related measures *(either same measure focus or target population)* or competing measures *(both the same measure focus and same target population*), list the NQF # and title of all related and/or competing measures:

5a. Harmonization

5a.1 If this measure has EITHER the same measure focus OR the same target population as NQF-endorsed measure(s):
Are the measure specifications completely harmonized?

5a.2 If the measure specifications are not completely harmonized, identify the differences, rationale, and impact on interpretability and data collection burden:

5b. Competing Measure(s)

5b.1 If this measure has both the same measure focus and the same target population as NQF-endorsed measure(s):
Describe why this measure is superior to competing measures *(e.g., a more valid or efficient way to measure quality);* OR

provide a rationale for the additive value of endorsing an additional measure. *(Provide analyses when possible)*:
n/a

CONTACT INFORMATION

Co.1 Measure Steward (Intellectual Property Owner): Society of Thoracic Surgeons, 633 N Saint Clair, Floor 23, Chicago, Illinois, 60611

Co.2 Point of Contact: Vadie, Reese, vreese@sts.org, 312-202-5856-

Co.3 Measure Developer if different from Measure Steward: Society of Thoracic Surgeons, 633 N Saint Clair, Floor 23, Chicago, Illinois, 60611

Co.4 Point of Contact: Vadie, Reese, vreese@sts.org, 312-202-5856-

Co.5 Submitter: Vadie, Reese, vreese@sts.org, 312-202-5856-, Society of Thoracic Surgeons

Co.6 Additional organizations that sponsored/participated in measure development:

Co.7 Public Contact: Vadie, Reese, vreese@sts.org, 312-202-5856-, Society of Thoracic Surgeons

ADDITIONAL INFORMATION

Workgroup/Expert Panel involved in measure development
Ad.1 Provide a list of sponsoring organizations and workgroup/panel members' names and organizations. Describe the members' role in measure development.
Members of the STS Task Force on Quality Initiatives provide surgical expertise as needed. The STS Workforce on National Databases meets at the STS Annual Meeting and reviews the measures on a yearly basis. Changes or updates to the measure will be at the recommendation of the Workforce.

Ad.2 If adapted, provide title of original measure, NQF # if endorsed, and measure steward. Briefly describe the reasons for adapting the original measure and any work with the original measure steward:

Measure Developer/Steward Updates and Ongoing Maintenance
Ad.3 Year the measure was first released: 2010
Ad.4 Month and Year of most recent revision: 12, 2011
Ad.5 What is your frequency for review/update of this measure? annually
Ad.6 When is the next scheduled review/update for this measure? 12, 2012

Ad.7 Copyright statement:

Ad.8 Disclaimers:

Ad.9 Additional Information/Comments:

Date of Submission (*MM/DD/YY*): 12/20/2011

1b.2. Summary of Measure Results Demonstrating Performance Gap *(Descriptive statistics for performance results for this measure - distribution of scores for measured entities by quartile/decile, mean, median, SD, min, max, etc.; variation or overall less than optimal performance across providers)*

Dates: January 1, 2008 – December 31, 2010

Data/Sample: The population included 22,677 records from 174 hospitals. Hospital-specific sample sizes ranged from 1 to 883 records per hospital.

Distribution of hospital-specific estimates of standardized incidence ratio (SIR) for mortality or morbidity:

Minimum	1st quartile	Median	Mean	3rd quartile	Maximum
0.52	0.86	1.04	1.06	1.20	2.18

1b.4. Summary of Measure Results on Disparities by Population Group *(Descriptive statistics for performance results for this measure by population group)*

Race	**Frequency**	**Percent in Population**	**Proportion with Mortality or Morbidity**	**95% Confidence Interval**
Caucasian	19871	87.6	0.087	0.083, 0.091
Black	1910	8.4	0.084	0.072, 0.098
Hispanic	110	0.5	0.100	0.051, 0.172
Asian	394	1.7	0.069	0.046, 0.098
Native American	31	0.1	0.032	0.001, 0.167
Hawaiian/Pacific	17	0.1	0.059	0.001, 0.287
Mixed	98	0.4	0.071	0.029, 0.142
Other	195	0.9	0.082	0.048, 0.130
Missing	51	0.2	0.098	0.033, 0.214

2a2.2. Analytic Methods *(Describe method of reliability testing and rationale)*

Bayesian hierarchical modeling was used to assess the statistical reliability of hospital-specific standardized incidence ratio (SIR) estimates derived from the January 1, 2008 – December 31, 2010 STS data. All hospitals regardless of sample size were included in the estimation of model parameters. Reliability measures were initially calculated including all the hospitals and were subsequently calculated in subsets of hospitals having at least 10, 20, 30, 50, 100, or 200 eligible cases.

Definition of Reliability

Reliability is conventionally defined as the proportion of variation in a performance measure that is due to true between-hospital differences (i.e., signal) as opposed to random statistical fluctuations (i.e., noise). A mathematically equivalent definition is the squared correlation between a measurement and the true value. Estimation of reliability for this particular performance measure is complicated by the fact that it is risk-adjusted using a Bayesian hierarchical model.

We used two different approaches to operationalize the above definitions of reliability in a Bayesian statistical framework. First, we calculated the proportion of variance in the Bayesian posterior distribution (i.e. the range of possible beliefs about the performance of different hospitals) that could be attributed to differences between hospitals as opposed to residual (within-hospital) uncertainty. A closely related reliability measure was defined as the squared correlation between each hospital's estimated performance measure (the estimated SIR) and the true value. The latter measure could not be calculated directly (because the "true" SIR values are unknown) but was estimated using Bayesian inference methods.

<u>Technical Details</u>

Let θ_j denote the true unknown SIR value for the *j*-th of *J* hospitals. Prior to estimating reliability, the numerical value of θ_j was estimated for each hospital under the model described by Kozower et al. (2008). Estimation was done using Markov Chain Monte Carlo (MCMC) simulations and involved the following steps:

1. First, for each *j*, we randomly generated a large number *N* of possible numerical values of θ_j by sampling from the Bayesian posterior probability distribution of θ_j. Let $\theta_j^{(i)}$ denote the *i*-th of these *N* randomly sampled numerical values for the *j*-th hospital.
2. Second, for each *j*, a Bayesian estimate $\hat{\theta}_j$ of θ_j was calculated as the arithmetic average of the randomly sampled values $\theta_j^{(1)}, \dots, \theta_j^{(N)}$; in other words $\hat{\theta}_j = \frac{1}{N}\sum_{i=1}^{N}\theta_j^{(i)}$.

Uncertainty about each θ_j is reflected in the variance of the randomly sampled values $\theta_j^{(1)}, \dots, \theta_j^{(N)}$ about their mean. Greater variance implies greater uncertainty (and hence less reliability). To operationalize the concept of reliability, let us define

$$\sigma_{total}^2 = \frac{1}{JN}\sum_{j=1}^{J}\sum_{i=1}^{N}\left(\theta_j^{(i)} - \bar{\theta}\right)^2 \quad \text{where} \quad \bar{\theta} = \frac{1}{JN}\sum_{j=1}^{J}\sum_{i=1}^{N}\theta_j^{(i)}$$

and note that

$$\sigma_{total}^2 = \sigma_{between}^2 + \sigma_{within}^2$$

where

$$\sigma_{between}^2 = N\sum_{j=1}^{J}\left(\hat{\theta}_j - \bar{\theta}\right)^2 \qquad \text{and} \qquad \sigma_{within}^2 = \sum_{j=1}^{J}\sum_{i=1}^{N}\left(\theta_j^{(i)} - \hat{\theta}_j\right)^2.$$

Our first measure of reliability was defined as

$$R^2 = \frac{\sigma_{between}^2}{\sigma_{between}^2 + \sigma_{within}^2}.$$

It may be interpreted as the proportion of posterior variance attributed to between-hospital variation as opposed to residual uncertainty.

Our second reliability measure was defined as the estimated squared correlation between the set of hospital-specific estimates $\hat{\theta}_1, \dots, \hat{\theta}_J$ and the corresponding unknown true values $\theta_1, \dots, \theta_J$. Let ρ^2 denote the <u>unknown true</u> squared correlation of interest and let $\hat{\rho}^2$ denote <u>an estimate</u> of this quantity. The estimate was calculated as

$$\hat{\rho}^2 = \frac{1}{N}\sum_{i=1}^{N}\rho_{(i)}^2$$

where

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

A 95% Bayesian probability interval (PrI) for ρ^2 was obtained calculating the 2.5th and 97.5th percentiles of the set of numbers $\rho^2_{(1)},\ldots,\rho^2_{(N)}$.

2a2.3. Testing Results *(Provide reliability statistics and assessment of adequacy in the context of norms for the test conducted)*

Using the methods described above, we obtained the estimates:

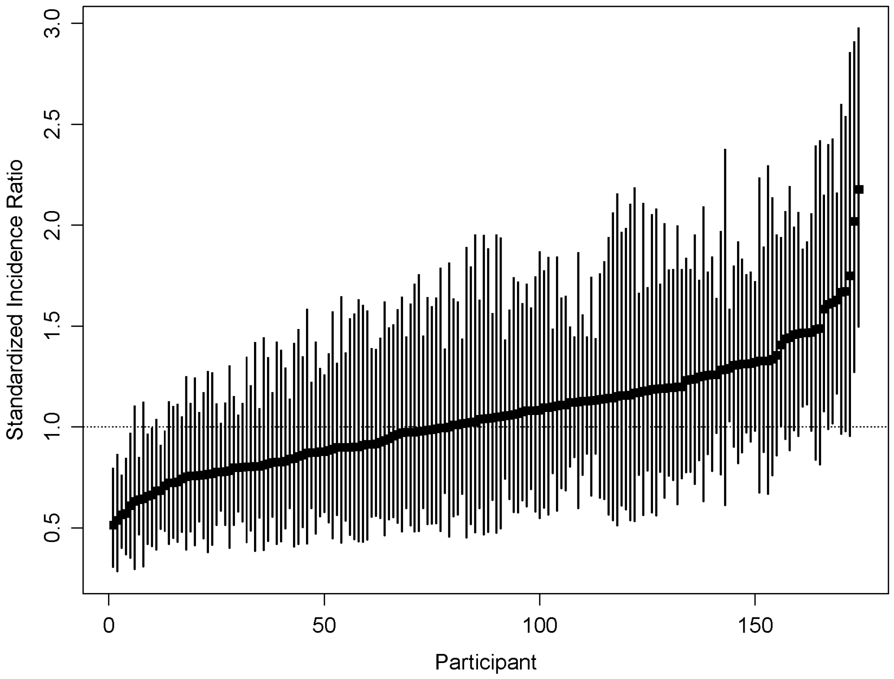
Sample Size Threshold	**Number of Hospitals Meeting Threshold**	**Number of Patients Included**	**Reliability R^2**	**Reliability $\hat{\rho}^2$ (95% PrI)**
≥ 1 case	174	22,677	0.50	0.51 (0.40, 0.61)
≥ 10 cases	160	22,605	0.54	0.55 (0.43, 0.65)
≥ 20 cases	155	22,535	0.55	0.56(0.45, 0.67)
≥ 30 cases	139	22,144	0.59	0.60 (0.48, 0.70)
≥ 50 cases	114	21,155	0.64	0.65 (0.54, 0.75)
≥ 100 cases	77	18,425	0.70	0.71 (0.59, 0.81)
≥ 200 cases	35	12,340	0.75	0.77 (0.62, 0.88)

With all hospitals included, the Bayesian estimate of the reliability measure ρ^2 is 0.51 and the 95% Bayesian probability interval for ρ^2 is (0.40, 0.61). Since the lower limit of this interval is 0.40, we may be highly confident (probability = 97.5%) that the true reliability ρ^2 is at least 0.40.

In summary, when estimated with 3 years of data, the proposed lung cancer morbidity and mortality measure is reliable enough to be useful in the context of feedback reporting for internal quality improvement initiatives. Public reporting may be considered for a subset of hospitals with a larger case volume.

2b5.3. Results *(Provide measure performance results/scores, e.g., distribution by quartile, mean, median, SD, etc.; identification of statistically significant and meaningfully differences in performance)*

Plot displays estimated SIR and corresponding 95% Bayesian probability interval for each of 174 hospitals. Hospitals are ordered according to the increasing SIR estimate. There are meaningful differences between the best performing and the worst performing hospitals.



Disparities in Care

2c.1. If measure is stratified for disparities, provide stratified results *(Scores by stratified categories/cohorts)*

Table summarizes proportions of patients with mortality or morbidity by race. Even though estimates of these proportions differ somewhat by race, the 95% confidence intervals overlap.

Race	Frequency	Percent in Population	Proportion with Mortality or Morbidity	95% Confidence Interval
Caucasian	19871	87.6	0.087	0.083, 0.091
Black	1910	8.4	0.084	0.072, 0.098
Hispanic	110	0.5	0.100	0.051, 0.172
Asian	394	1.7	0.069	0.046, 0.098
Native American	31	0.1	0.032	0.001, 0.167
Hawaiian/Pacific	17	0.1	0.059	0.001, 0.287
Mixed	98	0.4	0.071	0.029, 0.142
Other	195	0.9	0.082	0.048, 0.130
Missing	51	0.2	0.098	0.033, 0.214

STS Database Risk Models: Predictors of Mortality and Major Morbidity for Lung Cancer Resection

Benjamin D. Kozower, MD, MPH, Shubin Sheng, PhD, Sean M. O'Brien, PhD, Michael J. Liptay, MD, Christine L. Lau, MD, David R. Jones, MD, David M. Shahian, MD, and Cameron D. Wright, MD

Departments of Surgery & Public Health Sciences, University of Virginia Health System, Charlottesville, Virginia; Duke Clinical Research Institute, Duke University, Durham, North Carolina; Department of Cardiovascular and Thoracic Surgery, Rush University, Chicago, Illinois; and Division of Thoracic Surgery and Center for Quality and Safety, Massachusetts General Hospital, Boston, Massachusetts

Background. The aim of this study is to create models for perioperative risk of lung cancer resection using the STS GTDB (Society of Thoracic Surgeons General Thoracic Database).

Methods. The STS GTDB was queried for all patients treated with resection for primary lung cancer between January 1, 2002 and June 30, 2008. Three separate multivariable risk models were constructed (mortality, major morbidity, and composite mortality or major morbidity).

Results. There were 18,800 lung cancer resections performed at 111 participating centers. Perioperative mortality was 413 of 18,800 (2.2%). Composite major morbidity or mortality occurred in 1,612 patients (8.6%). Predictors of mortality include the following: pneumonectomy ($p < 0.001$), bilobectomy ($p < 0.001$), American Society of Anesthesiology rating ($p < 0.018$), Zubrod performance status ($p < 0.001$), renal dysfunction ($p = 0.001$), induction chemoradiation therapy ($p = 0.01$), steroids ($p = 0.002$), age ($p < 0.001$), urgent procedures ($p = 0.015$), male gender ($p = 0.013$), forced expiratory volume in one second ($p < 0.001$), and body mass index ($p = 0.015$).

Conclusions. Thoracic surgeons participating in the STS GTDB perform lung cancer resections with a low mortality and morbidity. The risk-adjustment models created have excellent performance characteristics and identify important predictors of mortality and major morbidity for lung cancer resections. These models may be used to inform clinical decisions and to compare risk-adjusted outcomes for quality improvement purposes.

(Ann Thorac Surg 2010;90:875–83)

The Society of Thoracic Surgeons (STS) has played a critical role for the past two decades in developing risk models to adjust cardiac surgery outcomes for preoperative patient characteristics and disease severity [1, 2]. These risk-adjusted outcomes have provided valuable information for research, patient counseling, quality assessment, and benchmark comparisons between providers and hospitals. The STS General Thoracic Surgery Database (GTSD) was established in 2002 to provide the same opportunity for general thoracic surgeons [3]. This database has been used successfully to investigate the surgical management of primary lung cancer and to identify predictors of prolonged length of stay after lobectomy [4, 5]. The database has also been used to identify predictors of major morbidity and mortality after esophagectomy for esophageal cancer [6].

The National Veterans Affairs Surgical Risk Study was the first large multicenter center study to evaluate risk-adjusted mortality rates for noncardiac surgery and create a prognostic model for major pulmonary resection [7, 8]. However, this model represents a limited patient population of male veterans which is difficult to generalize to the lung cancer population as a whole. The objectives of this study were to create models for the perioperative risk of lung cancer resection using the STS GTSD. We examined three different outcomes: mortality, major morbidity, and a composite outcome including mortality or major morbidity. We also evaluated these models to determine if they could measure variation in hospital performance. The clinical significance of these different models will have varying appeal to individual clinicians and patients. This is important because patient's attitudes on whether or not to accept the risk of surgery are guided by their risks of death and loss of independence [9, 10].

Patients and Methods

Society of Thoracic Surgeons Database

The STS GTSD was established as a voluntary initiative to support the continuous quality improvement efforts of surgeons and hospitals. Participating institutions receive twice-yearly feedback reports that describe each site's results in relation to other database participants. Although the database is not currently audited, all partici-

Accepted for publication March 19, 2010.

Presented at the Forty-sixth Annual Meeting of The Society of Thoracic Surgeons, Fort Lauderdale, FL, Jan 25–27, 2010. Winner of the J. Maxwell Chamberlain Memorial Award for General Thoracic Surgery.

Address correspondence to Dr Kozower, University of Virginia Health System, General Thoracic Surgery, PO Box 800679, Charlottesville, VA 22908-0679; e-mail: bdk8g@virginia.edu.

0003-4975/$36.00
doi:10.1016/j.athoracsur.2010.03.115

pants sign a contract that requires complete reporting of all cases and prohibits selective reporting. Details of the STS GTDB data collection instrument can be found on the STS website [11]. Participation in the STS GTDB requires initial institutional review board approval, but subsequent deidentified data analysis for quality improvement purposes does not. This study was approved by the University of Virginia Human Investigation Committee.

Patient Population

The STS GTDB was queried for all patients treated with resection for primary lung cancer between January 1, 2002 and June 30, 2008. Lung cancer resection included the following: lobectomy, sleeve lobectomy, bilobectomy, pneumonectomy, segmentectomy, and wedge resection. Data were excluded for benign disease in the pathologic staging, emergent operation, extrapleural pneumonectomy, missing age, missing gender, or missing mortality.

Outcome Definitions

Postoperative events were defined by the STS GTDB guidelines [11]. Perioperative mortality is defined as death during the same hospitalization as surgery or within 30 days of the procedure. We analyzed a composite morbidity-mortality outcome along with separate models for mortality and major morbidity. Selection of adverse outcome measures was based on clinical judgment, literature review, and preliminary data analysis. The composite outcome was defined as the presence of one or more of the following postoperative conditions: perioperative mortality, tracheostomy, reintubation, initial ventilatory support greater than 48 hours, adult respiratory distress syndrome, bronchopleural fistula, pulmonary embolus, pneumonia, bleeding requiring reoperation, and myocardial infarction.

Selection of Covariates

Model variables were identified by reviewing three versions of the STS data collection instrument (v1.3, v2.06, v2.07). Variables were selected a priori based on a combination of literature review and informal empirical analysis (Table 1). Because one purpose of the model was to adjust for case mix in making hospital comparisons, candidate predictor variables were limited to preoperative patient factors that were not directly modifiable by the surgeon or hospital. Missing predictor values were imputed using multiple imputation. A sensitivity analysis was conducted to compare results from two missing data software packages, Imputation and Variance Estimation Software (IVEWare) and MICE (multiple imputation using chained equations) [12, 13]. Results from these two different methods were almost identical and we report the MICE results.

Multivariable Analysis

Multivariable logistic regression was used to estimate the relationship between patient preoperative characteristics and the outcomes of mortality, major morbidity, and composite mortality or major morbidity. All covariates were retained in the model and were not added or removed based on a variable selection algorithm. Parameters of the logistic model were estimated using generalized estimating equations methodology to account for statistical dependence between outcomes of patients at the same hospital. Discrimination of the model was assessed by averaging the area under the receiver operating characteristic curve, also known as the C-statistic, from each of the ten imputation datasets. A split sample approach was used to calculate the C-statistic. The statistical significance of the difference in the observed to expected numbers of the three outcomes was assessed using the Hosmer-Lemeshow goodness-of-fit test.

Analysis of Hospital Performance Variation

To explore variation in hospital performance, the model described above for major morbidity or mortality was subsequently refit as a two-level hierarchic model with nesting of patients within participants. The hierarchic model included the same set of patient factors described above, plus a set of random hospital-specific effects. The hospital-specific effects are interpreted as reflecting underlying differences in performance that systematically increase or decrease risk of all patients at the same hospital. Performance variation was summarized by calculating the hospital-specific standardized incidence ratio (SIR) of mortality or major morbidity. The SIR is

Table 1. Baseline Characteristics

Variable	No. (% of All Patients)	Median
Total	18,800 (100)	
Male gender	9,212 (49.0)	
Race		
White	16,286 (86.6)	
Black	1,427 (7.6)	
Other	913 (4.9)	
Body mass index (kg/m^2)		26.4
CAD	3,935 (20.9)	
Diabetes	2,944 (15.6)	
Renal dysfunction	589 (3.1)	
Induction chemotherapy	684 (3.6)	
Induction chemoradiation therapy	1,698 (9.0)	
Recent cigarette use	5,036 (26.9)	
Steroids	727 (3.9)	
Thoracoscopy	6,947 (36.9)	
Thoracotomy	12,834 (68.3)	
Primary procedure		
Wedge resection	3,621 (19.6)	
Segmentectomy	818 (4.3)	
Lobectomy	12,313 (65.5)	
Sleeve lobectomy	269 (1.4)	
Bilobectomy	647 (3.4)	
Pneumonectomy	1,132 (6.0)	

CAD = coronary artery disease.

Table 2. Frequency of Complications

	All Patients		Patients With Mortality		Patients With Major Morbidity		Patients With Mortality or Major Morbidity	
Variable	n	% of All Patients	n	% of Subgroup	n	% of Subgroup	n	% of Subgroup
Pulmonary embolus	81	0.43	15	18.5	81	100.0	81	100.0
DVT	86	0.46	13	15.1	50	58.1	50	58.1
Tracheostomy	244	1.30	67	27.5	244	100.0	244	100.0
Atrial fibrillation	2,039	10.85	114	5.6	389	19.1	407	20.0
Myocardial infarction	67	0.36	21	31.3	67	100.0	67	100.0
Intraoperative blood transfusion	476	2.73	33	6.9	108	22.7	117	24.6
Postoperative blood transfusion	1106	6.34	127	11.5	424	38.3	443	40.1
RLN paralysis	66	0.35	2	3.0	14	21.2	14	21.2
Renal failure	257	1.37	88	34.2	137	53.3	138	53.7
Sepsis	153	0.81	81	52.9	120	78.4	128	83.7
Chylothorax	46	0.24	2	4.3	10	21.7	11	23.9

DVT = deep venous thrombosis; RLN = recurrent laryngeal nerve.

defined as the ratio of the participant's risk-adjusted rate divided by the risk-adjusted rate of a hypothetical "average" participant.

A SIR value greater than 1.0 implies that a participant's rate of mortality or major morbidity is higher than the rate that would be projected for an average participant that operated on the same case mix of patients. Uncertainty surrounding the estimated SIR was quantified by calculating Bayesian 95% probability intervals. Details of the hierarchic model, including the calculation and interpretation of SIRs and probability intervals, have been previously described [5, 6]. Analyses were performed using S-Plus 6 (Insightful Corp, Seattle, WA), SAS 9.1 (SAS Institute, Cary, NC), and WinBUGS 1.4.1 (Freeware, http://www.mrc-bsu.cam.ac.uk/bugs/welcome.shtml and Imperial College of Science, Technology and Medicine at St Mary's, London).

Results

Our initial query of the STS GTDB revealed 19,026 patients having surgery for lung cancer from 111 centers. Our final analysis included 18,800 of these patients, excluding records for benign disease for pathologic staging (n = 50), extrapleural pneumonectomy (n = 34), and for missing age, gender, or mortality (n = 142). Patient characteristics and some of the variables examined in the risk models are detailed in Table 1. The majority of patients were white (86%), had a good performance status (Zubrod score of 0 or 1, 87%) and had significant comorbidities (American Society of Anesthesiologists [ASA] rating of III or greater, 77%).

Perioperative mortality occurred in 413 patients (2.2%). Major morbidity occurred in 1,491 patients (7.9%). Median length of stay after lung cancer resection was 5 days and was significantly increased for patients having at least one major complication (5 days vs 12 days, $p <$ 0.001). The composite outcome of mortality or major morbidity occurred in 1,612 patients: perioperative mortality (n = 413), pneumonia (n = 722), reintubation (n = 654), tracheostomy (n = 244), adult respiratory distress syndrome (n = 220), initial ventilatory support greater than 48 hours (n = 176), bleeding requiring reoperation (n = 137), pulmonary embolus (n = 81), myocardial infarction (n = 67), and bronchopleural fistula (n = 60). In addition to the variables used to calculate major morbidity, postoperative complications shown in Table 2 include deep venous thrombosis, atrial fibrillation, intraoperative and postoperative blood transfusion, recurrent laryngeal nerve paralysis, renal failure, sepsis, and chylothorax.

The multivariable models illustrating the association

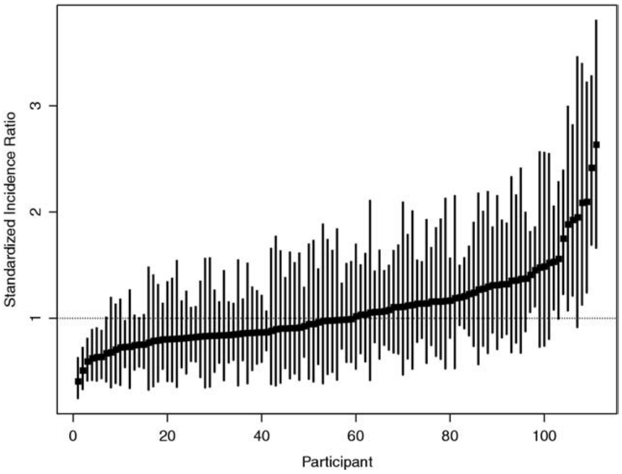


Fig 1. Hospital performance variation. The standardized incidence ratio of mortality or major morbidity after lung cancer resection among Society of Thoracic Surgeons participating sites are shown. The confidence intervals do not overlap for the best performing sites (on the far left) and the worse performers (on the far right).

Table 3. Predictors of Major Morbidity and Mortality

	Mortality Model		Major Morbidity Model		Composite Model (Mortality or Major Morbidity)	
Variable	OR (95% CI)	*p*	OR (95% CI)	*p*	OR (95% CI)	*p*
Age (10-year increase)	1.84 (1.58, 2.15)	<0.001	1.23 (1.15, 1.32)	<0.001	1.27 (0.18, 1.36)	<0.001
Male gender	1.36 (1.07, 1.73)	0.013	1.07 (0.95,1.21)	0.265	1.12 (0.01, 1.23)	0.031
Surgery year	0.99 (0.91, 1.07)	0.762	1.02 (0.94, 1.11)	0.682	0.99 (0.93, 1.05)	0.684
BMI (10 kg/m^2 increase)	0.74 (0.58, 0.94)	0.016	0.84 (0.73, 0.96)	0.011	0.83 (0.73, 0.94)	0.003
Hypertension	1.10 (0.85, 1.43)	0.459	1.10 (0.95, 1.28)	0.191	1.08 (0.93, 1.24)	0.312
Steroids	1.93 (1.26, 2.96)	0.003	1.60 (1.24,2.07)	<0.001	1.63 (1.31, 2.03)	<0.001
CHF	1.00 (0.62, 1.62)	0.994	1.52 (1.17,1.98)	0.002	1.43 (1.10, 1.86)	0.007
CAD	1.10 (0.79, 1.53)	0.562	1.18 (1.01,1.38)	0.035	1.18 (1.02, 1.38)	0.031
PVD	1.36 (0.97, 1.91)	0.070	1.02 (0.82,1.28)	0.830	1.10 (0.90, 1.34)	0.369
Thoracic reoperation	0.51 (0.25, 1.03)	0.061	1.24 (0.94, 1.65)	0.133	1.21 (0.92, 1.59)	0.169
Cerebrovascular disease	1.09 (0.78, 1.51)	0.612	1.10 (0.88, 1.38)	0.413	1.16 (0.97, 1.40)	0.102
Diabetes	1.15 (0.82, 1.59)	0.422	1.07 (0.92, 1.24)	0.402	1.11 (0.95, 1.31)	0.199
% FEV$_1$ (10% decrease)	1.11 (1.18, 1.04)	0.001	1.08 (1.11, 1.03)	<0.001	1.06 (1.10, 1.03)	<0.001
Urgent vs elective	1.71 (1.11, 2.64)	0.015	0.97 (0.74, 1.29)	0.854	1.15 (0.87, 1.53)	0.317
Induction chemotherapy alone	1.00 (0.54, 1.84)	0.996	0.85 (0.57, 1.28)	0.438	0.88 (0.61, 1.28)	0.514
Induction chemoradiation therapy	2.12 (1.20, 3.76)	0.010	1.99 (1.32, 3.02)	0.001	1.91 (1.29, 2.83)	0.001
Creatinine ≥ 2	2.48 (1.42, 4.31)	0.001	1.67 (1.17, 2.38)	0.005	1.78 (1.27, 2.49)	0.001
Dialysis	3.97 (1.48,10.64)	0.006	1.95 (1.04, 3.66)	0.038	2.21 (1.24, 3.95)	0.007
Recent cigarette use	1.03 (0.60, 1.74)	0.924	1.50 (1.19, 1.89)	<0.001	1.47 (1.17, 1.85)	0.001
Zubrod (vs 0)						
4	6.32 (2.73,14.63)	<0.001	4.67 (2.48, 8.80)	<0.001	5.46(2.91,10.22)	<0.001
3	3.05 (1.95, 4.77)	<0.001	2.54 (1.76, 3.67)	<0.001	2.74 (1.92, 3.89)	<0.001
2	1.76 (1.28, 2.42)	0.005	1.60 (1.26, 2.04)	<0.001	1.64 (1.31, 2.05)	<0.001
1	1.21 (0.96, 1.53)	0.102	1.17 (1.01, 1.37)	0.039	1.17 (1.01, 1.35)	0.031
ASA (vs 1)						
5	5.09 (1.89,13.76)	0.001	3.85 (2.50, 5.95)	<0.001	3.64 (2.35, 5.64)	<0.001
4	4.79 (1.98,11.54)	<0.001	1.97 (1.35, 2.88)	<0.001	2.03 (1.41, 2.92)	0.001
3	3.57 (1.48, 8.63)	0.005	1.26 (0.84, 1.90)	0.270	1.35 (0.91, 1.99)	0.134
2	2.12 (1.14, 3.95)	0.018	1.00 (0.74, 1.36)	0.974	1.06 (0.80, 1.42)	0.671
Thoracotomy vs VATS	1.25 (0.90, 1.73)	0.189	1.58 (1.29, 1.94)	<0.001	1.55 (1.27, 1.88)	<0.001
Pathologic stage (vs I)						
Stage II	0.92 (0.64, 1.33)	0.656	1.00 (0.85, 1.19)	0.961	0.99 (0.83, 1.17)	0.866
Stage III	1.32 (0.82, 2.12)	0.256	1.15 (0.89, 1.49)	0.278	1.24 (0.96, 1.60)	0.094
Stage IV	2.02 (1.18, 3.47)	0.011	0.94 (0.61, 1.46)	0.786	1.09 (0.75, 1.59)	0.636
Procedure (vs wedge resection)						
Segmentectomy	1.25 (0.75, 2.10)	0.391	2.13 (1.52, 2.99)	<0.001	2.07 (1.55, 2.76)	<0.001
Lobectomy	1.03 (0.74, 1.44)	0.849	1.93 (1.56, 2.38)	<0.001	1.69 (1.39, 2.06)	0.005
Sleeve lobectomy	1.59 (0.45, 5.60)	0.472	2.00 (1.15, 3.47)	0.014	2.10 (1.24, 3.54)	<0.001
Bilobectomy	2.61 (1.71, 4.00)	<0.001	3.22 (2.39, 4.34)	<0.001	3.03 (2.35, 3.92)	<0.001
Pneumonectomy	3.91 (2.46, 6.22)	<0.001	2.54 (1.82, 3.54)	<0.001	2.54 (1.85, 3.49)	<0.001
C statistic	0.77		0.69		0.69	
Hosmer-Lemeshow *p* Value	0.17		0.46		0.47	

ASA = American Society of Anesthesiologists; BMI = body mass index; CAD = coronary artery disease; CHF = congestive heart failure; CI = confidence interval; FEV$_1$ = forced expiratory volume in the first second of expiration; OR = odds ratio; PVD = peripheral vascular disease; VATS = video-assisted thoracic surgery.

between preoperative patient characteristics and the endpoints of mortality, major morbidity, and the composite outcome are summarized in Table 3. The bootstrap-adjusted C-statistics for the models are 0.77, 0.69, and 0.69, respectively. The Hosmer-Lemeshow goodness-of-fit *p* values for all three models demonstrate that the model estimates fit the data at an acceptable level (Table 3). Poor performance status (high Zubrod score) and poor physical status of patients before surgery (high ASA rating) have large impacts on mortality and morbidity. Variables such as thoracic reoperation and induction chemotherapy, which many consider to increase perioperative risk, were not associated with mortality or morbidity in any of the three models. However, induction chemoradiation therapy was a predictor of both mortality and major morbidity (Table 3). Of note, compared with a nonanatomic wedge resection, segmentectomy and lobectomy increased the risk of major complications but not the risk of mortality.

Figure 1 illustrates the standardized incidence ratio for the composite outcome of mortality or major morbidity for the 111 hospitals. This is interpreted as the rate of mortality or major morbidity at a particular hospital compared with the projected rate for an average participant that treated the same case mix of patients. The probability intervals of some of the best performing hospitals (on the left side) do not overlap with some of the hospitals with worse outcomes (on the right side). This indicates that the model facilitates a meaningful comparison between hospitals with a significant difference between some of the best and worst performers.

Comment

The STS GTDB has grown considerably over the past decade and provides the thoracic surgery community with this opportunity to evaluate 18,800 lung cancer resection patients. Hospitals participating in the STS GTDB perform lung cancer resections with low mortality (2.2%) and major morbidity (7.9%). These excellent outcomes are similar to the mortality reported by the prospectively conducted American College of Surgeons Z0030 trial [14] and considerably lower than the 5.2% mortality from the National Veterans Affairs Surgical Quality Improvement Program [8].

We identified 12 risk factors for mortality after lung cancer resection (Table 3). The Zubrod performance status and ASA rating are two of the strongest predictors of mortality. Zubrod performance status has been used by cooperative oncology trials for the past two decades and is an excellent predictor of survival [15]. The ASA score is a validated marker of medical comorbidity, a good index of case complexity, and a known predictor of perioperative mortality [16, 17]. Renal dysfunction, both dialysis and an elevated creatinine, is often accompanied by cardiovascular disease and is a known risk factor for morbidity after lung resection [5, 18].

Induction chemoradiation therapy, but not induction chemotherapy alone, was an independent predictor of mortality and major morbidity. Induction chemoradiation therapy patients had double the risk of major morbidity, which contributes to their increased risk of perioperative mortality. This finding is consistent with a previous STS GTDB study showing that induction therapy was also a risk factor for prolonged length of stay after lobectomy [5]. Currently, there is no clear consensus for the management of operable N2 disease (ipsilateral mediastinal lymph node involvement). Patients may be treated with surgery followed by adjuvant therapy, induction therapy followed by surgery, or definitive chemoradiation therapy alone. Some series of induction therapy for advanced stage lung cancer do not show an increased risk of perioperative mortality [19–21]. In addition, the recent STS GTDB model for esophageal cancer found no significant increased risk for patients having induction therapy [6]. The increased risk of induction chemoradiation therapy found in this study is an interesting finding as this is by far the largest study to date. It will be important to determine if this finding holds true in prospective trials and to determine whether there is a long-term survival benefit of induction therapy that offsets the increased perioperative risk.

Steroids are an important predictor of morbidity and mortality after thoracic surgery [6]. We found that steroid use almost doubled the risk of mortality (Table 3; OR = 1.93) and significantly increased the risk of major morbidity (Table 3; OR = 1.60). The forced expiratory volume in the first second of expiration (FEV_1) modeled as a 10% decrease in the percent predicted FEV_1 was also an independent predictor of increased mortality in the model but the effect size was relatively small (Table 3; OR = 1.11). Intuitively, the importance of a 10% decrease in FEV_1 would be greater for a starting FEV_1 of 45% than 80%. Further investigation looking at the nonlinear contribution of pulmonary function as well as other pulmonary function measurements not modeled in this study will be important.

Increasing body mass index (BMI) decreased the risk of mortality in our model. There were 8.5% of patients in the study with a BMI equal to 35 or greater and their mortality was only 1.6%. Previous researchers have hypothesized that obesity would increase postoperative complications but found no significant impact after 500 anatomic resections for lung cancer [22]. Our finding is partially explained by an increased rate of major morbidity and mortality in the low BMI group; BMI 25 or less (Table 1). Unfortunately, albumin was not modeled due to the large amount of missing data (38%) so an association cannot be made between poor nutritional status and low BMI.

Male gender was an independent predictor for morbidity and mortality in our model. This finding is difficult to explain but has been demonstrated in other lung resection models from the United States and Europe [5, 23]. A recent Japanese study of 2,770 patients with resected lung cancer found that women also had better 5-year survival than men [24].

The increased risk of mortality after pneumonectomy is well documented. This study confirms the considerable increased risk for pneumonectomy and bilobectomy (OR

= 3.91 and 2.61, respectively). Compared with a nonanatomic wedge resection, lobectomy and segmentectomy were not predictors of mortality. However, they were predictors of major morbidity. This finding is likely explained by the increased complexity and operative time of an anatomic resection. Although lobectomy has been the standard of care for early stage lung cancer, sublobar resection may provide a similar oncologic result for node negative, T1a tumors, and is currently being prospectively evaluated (CALGB 140503) [25].

Surgical approach, thoracoscopy versus thoracotomy, was not a predictor of mortality in this study. However, a thoracoscopic approach reduced the chance of major morbidity and the composite outcome. This finding supports recent single-center studies suggesting that minimally invasive approaches reduce perioperative complications after lung resection [26, 27].

Recent tobacco use was a predictor of major morbidity but not mortality. Eighty-five percent of patients were former or current smokers and almost one third of patients were recent smokers. Their increased risk of complications translates into longer hospital stays and increased cost. Some surgeons may hesitate to council their patients on tobacco cessation within 2 months of surgery for fear of a paradoxical increased risk of morbidity. However, there does not appear to be an increase in pulmonary complications when smokers quit within 2 months of surgery [28]. The study by Barrera and colleagues [28] is the largest prospective trial focusing specifically on this question and should help thoracic surgeons take advantage of an important teachable moment to help our patients and their family members quit smoking [29, 30].

Figure 1 illustrates that the model of composite mortality or major morbidity facilitates a meaningful comparison of quality between hospitals. The majority of the 111 hospitals perform in a similar fashion, as evidenced by a standardized incidence ratio near 1 with confidence intervals that overlap. However, there are significant differences between some of the best (left side) and worst (right side) performers (Fig 1).

There are several limitations of our report. First, the STS GTDB has a selection bias as participants are likely to be academic general thoracic surgeons. The specialty training of these surgeons has been shown to improve outcomes and their results may not be generalizable to cardiac or general surgeons performing limited amounts of general thoracic surgery. Second, the database is currently not audited for data quality. The STS General Thoracic Surgery Database Taskforce is aware of this problem and audits are scheduled to begin in 2010. There are no rewards for any potential "gaming," so it is unlikely to be a systemic issue. A third limitation of this study is missing data. Rather than eliminating records with missing data, we used multiple imputation techniques to create our models. However, this was problematic for variables such as clinical staging and diffusion capacity as both variables were missing in almost 40% of patient records. Given the large amount of missing data for these variables, we did not include them in the models. We examined pathologic stage as a surrogate for clinical stage knowing that it is not known preoperatively, but it was not a predictor of mortality or morbidity except for stage IV versus stage I patients. Diffusion capacity is a well-known predictor of outcome after lung resection and its omission likely decreases model performance [31–33]. Fourth, the STS GTDB is currently limited to 30-day follow-up. This does not have a major impact on the creation of perioperative models but it prevents research looking at longer term outcomes. To address this issue, the STS Workforce on National Databases is in the process of linking the STS databases with survival databases.

In conclusion, thoracic surgeons participating in the STS GTDB perform lung cancer resections with a low mortality and morbidity. Predictors of mortality include the following: pneumonectomy, bilobectomy, American Society of Anesthesiology rating, Zubrod performance status, renal dysfunction, induction chemoradiation therapy, steroids, age, urgent procedures, male gender, forced expiratory volume in one second, and body mass index. These models have excellent performance characteristics and will help surgeons and patients estimate perioperative risk and provide risk-adjusted outcomes for quality improvement.

Dr Kozower is supported by a grant from the Agency for Healthcare Research and Quality: K08 - HS18049.

References

1. Shahian DM, Edwards FH. The Society of Thoracic Surgeons 2008 cardiac surgery risk models: introduction. Ann Thorac Surg 2009;88(1 Suppl):S1.
2. Kozower BD, Ailawadi G, Jones DR, et al. Predicted risk of mortality models: Surgeons need to understand limitations of the University Health System Consortium models. J Am Coll Surg 2009;209:551–6.
3. Wright CD, Edwards FH. Society of Thoracic Surgeons General Thoracic Surgery database. Ann Thorac Surg 2007; 833:893–4.
4. Boffa DJ, Allen MS, Grab JD, Gaissert HA, Harpole DH, Wright CD. Data from the Society of Thoracic Surgeons General Thoracic Surgery database: the surgical management of primary lung tumors. J Thorac Cardiovasc Surg 2008;135:247–54.
5. Wright CD, Gaissert HA, Grab JD, O'Brien SM, Peterson ED, Allen MS. Predictors of prolonged length of stay after lobectomy for lung cancer: A Society Of Thoracic Surgeons General Thoracic Surgery Database risk-adjustment model. Ann Thorac Surg 2008;85:1857–65.
6. Wright CD, Kucharczuk JC, O'Brien SM, Grab JD, Allen MS. Society of Thoracic Surgeons General Thoracic Surgery Database. Predictors of major morbidity and mortality after esophagectomy for esophageal cancer: a Society of Thoracic Surgeons General Thoracic Surgery Database risk adjustment model. J Thorac Cardiovasc Surg 2009;137:587–95.
7. Khuri SF, Daley J, Henderson W, et al. Risk adjustment of the postoperative mortality rate for the comparative assessment of the quality of surgical care: Results of the national veterans affairs surgical risk study. J Am Coll Surg 1997;185: 315–27.
8. Harpole DH Jr, DeCamp MM Jr, Daley J, et al. Prognostic models of thirty-day mortality and morbidity after major

pulmonary resection. J Thorac Cardiovasc Surg 1999;117: 969–79.

9. Cykert S. Risk acceptance and risk aversion: patients' perspectives on lung surgery. Thorac Surg Clin 2004;14:287–93.
10. Cykert S, Kissling G, Hansen CJ. Patient preferences regarding possible outcomes of lung resection: what outcomes should preoperative evaluations target? Chest 2000;117: 1551–9.
11. Society of Thoracic Surgeons 2010. Available at: http://www.sts.org/sections/stsnationaldatabase/. Accessed January 4, 2010.
12. Taylor JM, Cooper KL, Wei JT, Sarma AV, Raghunathan TE, Heeringa SG. Use of multiple imputation to correct for nonresponse bias in a survey of urologic symptoms among African-American men. Am J Epidemiol 2002;156:774–82.
13. Giorgi R, Belot A, Gaudart J, Launoy G, French Network of Cancer Registries FRANCIM. The performance of multiple imputation for missing covariate data within the context of regression relative survival analysis. Stat Med 2008;27:6310–31.
14. Allen MS, Darling GE, Pechet TT, et al. Morbidity and mortality of major pulmonary resections in patients with early-stage lung cancer: Initial results of the randomized, prospective ACOSOG Z0030 trial. Ann Thorac Surg 2006;81: 1013–9.
15. Buccheri G, Ferrigno D, Tamburini M. Karnofsky and ECOG performance status scoring in lung cancer: a prospective, longitudinal study of 536 patients from a single institution. Eur J Cancer 1996;32A:1135–41.
16. Rantner B, Eckstein HH, Ringleb P, et al. American Society of Anesthesiology and Rankin as predictive parameters for the outcome of carotid endarterectomy within 28 days after an ischemic stroke. J Stroke Cerebrovasc Dis 2006;15:114–20.
17. Söderqvist A, Ekström W, Ponzer S, et al. Prediction of mortality in elderly patients with hip fractures: A two-year prospective study of 1,944 patients. Gerontology 2009;55: 496–504.
18. Obuchi T, Hamanaka W, Yoshida Y, et al. Clinical outcome after pulmonary resection for lung cancer patients on hemodialysis. Ann Thorac Surg 2009;88:1745–8.
19. Cerfolio RJ, Bryant AS, Jones VL, Cerfolio RM. Pulmonary resection after concurrent chemotherapy and high dose (60Gy) radiation for non-small cell lung cancer is safe and may provide increased survival. Eur J Cardiothorac Surg 2009;35:718–23.
20. Robinson LA, Ruckdeschel JC, Wagner H Jr, Stevens CW; American College of Chest Physicians. Treatment of non-small cell lung cancer-stage IIIA: ACCP evidence-based clinical practice guidelines (2nd edition). Chest 2007;132(3 Suppl):243S–65S.
21. Gaissert HA, Keum DY, Wright CD, et al. POINT: Operative risk of pneumonectomy--influence of preoperative induction therapy. J Thorac Cardiovasc Surg 2009;138:289–94.
22. Smith PW, Wang H, Gazoni LM, Shen KR, Daniel TM, Jones DR. Obesity does not increase complications after anatomic resection for non-small cell lung cancer. Ann Thorac Surg 2007;84:1098–105.
23. Berrisford R, Brunelli A, Rocco G, et al. The European Thoracic Surgery Database project: modelling the risk of in-hospital death following lung resection. Eur J Cardiothorac Surg 2005;28:306–11.
24. Chang JW, Asamura H, Kawachi R, Watanabe S. Gender difference in survival of resected non-small cell lung cancer: histology-related phenomenon? J Thorac Cardiovasc Surg 2009;137:807–12.
25. National cancer institute, clinical trials. Available at: http://www.cancer.gov/search/ViewClinicalTrials.aspx?cdrid=555324&version=HealthProfessional&protocolsearchid=7201104. Accessed January 11, 2010.
26. Atkins BZ, Harpole DH, Jr, Mangum JH, Toloza EM, D'Amico TA, Burfeind WR Jr. Pulmonary segmentectomy by thoracotomy or thoracoscopy: reduced hospital length of stay with a minimally-invasive approach. Ann Thorac Surg 2007;84:1107–12.
27. Villamizar NR, Darrabie MD, Burfeind WR, et al. Thoracoscopic lobectomy is associated with lower morbidity compared with thoracotomy. J Thorac Cardiovasc Surg 2009;138: 419–25.
28. Barrera R, Shi W, Amar D, et al. Smoking and timing of cessation: impact on pulmonary complications after thoracotomy. Chest 2005;127:1977–83.
29. Murin S. Smoking cessation before lung resection. Chest 2005;127:1873–5.
30. Kozower BD, Lau CL, Phillips JV, Burks SG, Jones DR, Stukenborg GJ. A thoracic surgeon directed tobacco cessation intervention. Ann Thorac Surg 2010;89:926–30.
31. Takeda S, Funakoshi Y, Kadota Y, et al. Fall in diffusing capacity associated with induction therapy for lung cancer: A predictor of postoperative complication? Ann Thorac Surg 2006;82:232–6.
32. Brunelli A, Refai MA, Salati M, Sabbatini A, Morgan-Hughes NJ, Rocco G. Carbon monoxide lung diffusion capacity improves risk stratification in patients without airflow limitation: Evidence for systematic measurement before lung resection. Eur J Cardiothorac Surg 2006;29:567–70.
33. Ferguson MK, Gaissert HA, Grab JD, Sheng S. Pulmonary complications after lung resection in the absence of chronic obstructive pulmonary disease: the predictive role of diffusing capacity. J Thorac Cardiovasc Surg 2009;138: 1297–302.

DISCUSSION

DR BILL PUTNAM (Nashville, TN): President Murray, Dr Wood, members of the Society and guests. It is an honor to discuss the J. Maxwell Chamberlain Paper for General Thoracic Surgery. Congratulations to the authors on a well-analyzed, articulate manuscript and presentation discussing the most important outcomes of lung cancer surgery. The Society of Thoracic Surgeons created the General Thoracic Surgery Database in 2002. Your focus on outcomes and quality rather than specialty or volume reflect the coming changes in health care. The academic focus of those investigators with a Master's degree in public health or a Master of Science in clinical investigations will create the evidence base for changes in our medical practice.

The careful analysis describes the current state of lung cancer surgery in a multicenter cohort of nearly 20,000 patients. This descriptive study noted a perioperative mortality of 2.2%, only slightly higher than the American College of Surgeons Oncology Group Z0030 clinical trial at 1.37%.

The predictors of mortality were those expected, including decreased physiologic status, poor performance, and low ASA (American Society of Anesthesiologists) rating. Use of these parameters allowed risk models to be constructed for either morbidity, mortality, or both. Missing data was noted. We do need better ways to have complete data. The authors created a hospital-specific measure to calculate an observed to expected rate of mortality or major morbidity. Good performing hospitals and poor performing hospitals were identified. Appropriately, smoking cessation is recommended.

This paper adds to the growing critical need for cardiothoracic surgeons to participate in the general thoracic database as a tangible means of quality improvement. I have three questions.

First, diversity. You described that the majority, 87%, of patients undergoing lung resection were white. The changing demographics in this country require thoughtful attention to access to care and analysis of outcomes. How do we improve access, data collection, and outcomes for all of our patients?

The second question, selection of patients. As cardiothoracic surgeons, we select patients based upon a combined personal experience model and sometimes a seat-of-the-pants physiologic model as well, usually based on patient performance state and pulmonary function. Is a personal risk profile available for the individual patient and the cardiothoracic surgeon to guide therapeutic recommendations based upon a validated model of predicted morbidity and mortality?

And thirdly, application of these exciting results. Within the general thoracic database you identified hospitals who were best performers. We need to define those characteristics and apply the best practices. Let's steal shamelessly from one another and improve care. How do you envision applying the results from the best hospitals to reduce morbidity and mortality in the others? How do we provide feedback to improve care?

I am very pleased to see this analysis of our national results from multiple centers. I believe the future looms just a bit brighter for our cardiothoracic surgeons and our surgeon investigators as we improve the outcomes of care for our patients with lung cancer.

Thank you for your attention.

DR KOZOWER: Thank you for your kind comments and your thoughtful questions. Your first question refers to diversity of the database. This may be an issue for how generalizable the models are and I agree with your comment that we need to continue to increase participation in the STS (Society of Thoracic Surgeons) database, particularly in underserved areas. That is probably the best way to increase population diversity.

Your second question addresses patient selection preoperatively. Although the cardiac surgery database has been used to develop a risk calculator, we are not quite there yet with the general thoracic database. However, as our data becomes more complete and participation increases, that is a very worthwhile goal.

Your third question pertains to the application of these models. Over 10% of hospitals in this analysis had statistically significant differences in performance. This is extremely important as it indicates that the combined mortality and major morbidity model can distinguish between the best and worst performers. I think we can learn something from our cardiac surgery colleagues. The VCSQI is the Virginia Cardiac Surgery Quality Initiative, a voluntary group of Virginia hospitals performing over 99% of the open-heart procedures in the state. The VCSQI exchanges data and implements protocols to improve outcomes. Importantly, they continue to study the impact of their interventions. STS Database participants could have a similar arrangement and we need to capitalize on this opportunity.

DR BRYAN FITCH MEYERS (St. Louis, MO): With regard to applying these findings, you don't know who the centers are on your graph, it is blinded to you, and I, like other STS Thoracic Database participants, am assuming that I am in that group that have hazards below the average, but I don't know that. How do you work to get the information out? How does the STS, how do the authors of this paper allow individual sites to break the code and find out where they fit in your risk-adjusted model?

DR KOZOWER: Dr Meyers, thank you for your question and for serving as a role model during my training and my first few years as an academic surgeon. Although the results of the entire graph appear blinded, each center knows their specific code so they can see how they compare to the group. Data is analyzed biannually so centers have a relatively real time look at how they are doing.

DR MEYERS: Is that in your analysis as well?

DR KOZOWER: This model will likely be used for the next data harvest.

DR MEYERS: Because the information we can get from the STS is not as risk adjusted as what you have created here.

DR KOZOWER: Correct, the current lung resection model is the model for prolonged length of stay. This model was developed to improve upon the previous model as the number of participating centers and patients has increased.

DR JOHN R. BENFIELD (Los Angeles, CA): In 1995, when I had the privilege of giving the presidential address to the STS, I addressed the issue that you have just brought to our attention. One of the few data points we had was Carolyn Reed's information from rural areas of the Southeast United States. Thanks to Professor Joachim Hasse of Freiburg, Germany, and others in Europe, we knew that about 80% of general thoracic surgery in Europe was being done by general surgeons. In Northern California, where I was working in 1995, much (perhaps most) of general thoracic surgery was being done by general surgeons.

My address was entitled "Metamorphosis" (Ann Thorac Surg 1996; 61:1045–1050). My proposal was controversial, and it generated no traction. This was probably because organized thoracic surgery did not fully accept the reality of so much thoracic surgery being done by general surgeons. In 2008 this phenomenon is still happening. Many patients are not getting the best available care that they need and deserve.

I take this opportunity to reiterate my proposal of 1995. I suggest that the STS grapple with this issue constructively and aggressively. We should offer to evaluate the work of general surgeons who are doing thoracic surgery, seeking to identify those whose current work is acceptable, despite their lack of background and training that meets our standards. Those general surgeons who are doing acceptable thoracic surgery, and know their limitations, should then be taken into our fold. They could then come and participate in our continuing education meetings. We should suggest a minimum of continuing education in thoracic surgery. We should create a grandfather clause, with a reasonable endpoint. We should urge CMS (Centers for Medicare & Medicaid Services) and other insurance carriers thereafter no longer to pay for thoracic surgery done by incompletely educated surgeons. Thereby we would elevate the standard of thoracic surgery in the US. I would appreciate your comments.

DR KOZOWER: That is a very good question and deserves consideration. We heard a great talk yesterday from Dr Wood looking at that same issue. Some of the participants in the database are likely to be performing cardiac surgery. The database has also been opened up to general surgeons, but their participation is still quite small.

DR OZ SHAPIRA (Jerusalem, Israel): This is kind of a challenging question. The STS was a pioneer in establishing the cardiac and then the thoracic and congenital databases. The STS encourages increased participation to increase the volume of the cases and therefore, the strength of its databases, yet there is continued refusal to add international sites. The STS member roster includes many international members practicing in international centers around the world who are very interested in participating in the STS database. I think it would greatly enhance the strength and value of the database. What is your thought about that?

DR KOZOWER: I think it is an excellent question. It is something that needs to be considered, especially when we see that one of the issues with our database is that it is fairly homogeneous with 87% Caucasians. I think it would be an excellent way for us to increase our population diversity. However, I am not an official spokesperson for the STS and don't know all of the details involved with such a policy.

DR DOUGLAS E. WOOD (Seattle, WA): And I will just comment that this is well recognized within the database workforce, and the fact is there is a task force in the database addressing international relations and involvement. So that is being actively addressed.

DR RICHARD J. SHEMIN (Los Angeles, CA): Do you have any idea of how many of the surgeons that actually participated are cardiac surgeons? We know from our workforce studies that probably 70% of surgeons that identify themselves as adult cardiac surgeons do some volume of general thoracic surgery. So is there any way to tease out that component from your data set?

DR KOZOWER: I don't know the exact number of participants that are also performing cardiac surgery. The majority are general thoracic surgeons, but it is obviously not all. That will be an important question for the database taskforce and will be important to answer Dr Benfield's question as well.

DR SHEMIN: Obviously, as thoracic surgeons, we have often tried to focus on the general surgeon performing thoracic surgery, we also have to be sure that the quality of the cardiac surgeon performing general thoracic surgery is equivalent to the full-time thoracic surgeon.