**National Quality Forum—Measure Testing (subcriteria 2a2, 2b1-2b6)**

**Measure Number** (*if previously endorsed*)**:** 1790

**Measure Title**: Risk-Adjusted Morbidity and Mortality for Lung Resection for Lung Cancer

**Date of Submission**: 11/15/2017

**Type of Measure:**

|  |  |
| --- | --- |
| Outcome (*including PRO-PM*) | Composite – ***STOP – use composite testing form*** |
| Intermediate Clinical Outcome | Cost/resource |
| Process *(including Appropriate Use)* | Efficiency |
| Structure |  |

|  |
| --- |
| **Instructions**   * Measures must be tested for all the data sources and levels of analyses that are specified. ***If there is more than one set of data specifications or more than one level of analysis, contact NQF staff*** about how to present all the testing information in one form. * **For all measures, sections 1, 2a2, 2b1, 2b2, and 2b4 must be completed.** * **For outcome and resource use measures**, section **2b3** also must be completed. * If specified for **multiple data sources/sets of specificaitons** (e.g., claims and EHRs), section **2b5** also must be completed. * Respond to all questions as instructed with answers immediately following the question. All information on testing to demonstrate meeting the subcriteria for reliability (2a2) and validity (2b1-2b6) must be in this form. An appendix for *supplemental* materials may be submitted, but there is no guarantee it will be reviewed. * If you are unable to check a box, please highlight or shade the box for your response. * Maximum of 25 pages (*incuding questions/instructions;* minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.*** * Contact NQF staff regarding questions. Check for resources at [Submitting Standards webpage](http://www.qualityforum.org/Measuring_Performance/Submitting_Standards.aspx). * For information on the most updated guidance on how to address social risk factors variables and testing in this form refer to the release notes for version 7.1 of the Measure Testing Attachment. |

|  |
| --- |
| **Note:** The information provided in this form is intended to aid the Standing Committee and other stakeholders in understanding to what degree the testing results for this measure meet NQF’s evaluation criteria for testing.  **2a2.** **Reliability testing** [**10**](#Note10) demonstrates the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise. For **instrument-based measures** (including PRO-PMs) **and composite performance measures**, reliability should be demonstrated for the computed performance score.  **2b1.** **Validity testing** [**11**](#Note11) demonstrates that the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality. For **instrument-based measures (including PRO-PMs) and composite performance measures**, validity should be demonstrated for the computed performance score.    **2b2.** **Exclusions** are supported by the clinical evidence and are of sufficient frequency to warrant inclusion in the specifications of the measure; [**12**](#Note12)  **AND**  If patient preference (e.g., informed decisionmaking) is a basis for exclusion, there must be evidence that the exclusion impacts performance on the measure; in such cases, the measure must be specified so that the information about patient preference and the effect on the measure is transparent (e.g., numerator category computed separately, denominator exclusion category computed separately). [**13**](#Note13)  **2b3.** **For outcome measures and other measures when indicated** (e.g., resource use):   * **an evidence-based risk-adjustment strategy** (e.g., risk models, risk stratification) is specified; is based on patient factors (including clinical and social risk factors) that influence the measured outcome and are present at start of care; [**14**](#Note14)**,**[**15**](#Note15) and has demonstrated adequate discrimination and calibration   **OR**   * rationale/data support no risk adjustment/ stratification.   **2b4.** Data analysis of computed measure scores demonstrates that methods for scoring and analysis of the specified measure allow for **identification of statistically significant and practically/clinically meaningful** [**16**](#Note16) **differences in performance**;  **OR**  there is evidence of overall less-than-optimal performance.  **2b5.** **If multiple data sources/methods are specified, there is demonstration they produce comparable results**.  **2b6.** Analyses identify the extent and distribution of **missing data** (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias.  **Notes**  **10.** Reliability testing applies to both the data elements and computed measure score. Examples of reliability testing for data elements include, but are not limited to: inter-rater/abstractor or intra-rater/abstractor studies; internal consistency for multi-item scales; test-retest for survey items. Reliability testing of the measure score addresses precision of measurement (e.g., signal-to-noise).  **11.** Validity testing applies to both the data elements and computed measure score. Validity testing of data elements typically analyzes agreement with another authoritative source of the same information. Examples of validity testing of the measure score include, but are not limited to: testing hypotheses that the measures scores indicate quality of care, e.g., measure scores are different for groups known to have differences in quality assessed by another valid quality measure or method; correlation of measure scores with another valid indicator of quality for the specific topic; or relationship to conceptually related measures (e.g., scores on process measures to scores on outcome measures). Face validity of the measure score as a quality indicator may be adequate if accomplished through a systematic and transparent process, by identified experts, and explicitly addresses whether performance scores resulting from the measure as specified can be used to distinguish good from poor quality. The degree of consensus and any areas of disagreement must be provided/discussed.  **12.** Examples of evidence that an exclusion distorts measure results include, but are not limited to: frequency of occurrence, variability of exclusions across providers, and sensitivity analyses with and without the exclusion.  **13.** Patient preference is not a clinical exception to eligibility and can be influenced by provider interventions.  **14.** Risk factors that influence outcomes should not be specified as exclusions.  **15.** With large enough sample sizes, small differences that are statistically significant may or may not be practically or clinically meaningful. The substantive question may be, for example, whether a statistically significant difference of one percentage point in the percentage of patients who received smoking cessation counseling (e.g., 74 percent v. 75 percent) is clinically meaningful; or whether a statistically significant difference of $25 in cost for an episode of care (e.g., $5,000 v. $5,025) is practically meaningful. Measures with overall less-than-optimal performance may not demonstrate much variability across providers. |

**1. DATA/SAMPLE USED FOR ALL TESTING OF THIS MEASURE**

*Often the same data are used for all aspects of measure testing. In an effort to eliminate duplication, the first five questions apply to all measure testing. If there are differences by aspect of testing,(e.g., reliability vs. validity) be sure to indicate the specific differences in question 1.7.*

**1.1. What type of data was used for testing**? (*Check all the sources of data identified in the measure specifications and data used for testing the measure*. *Testing must be provided for all the sources of data specified and intended for measure implementation.* ***If different data sources are used for the numerator and denominator, indicate N [numerator] or D [denominator] after the checkbox.***)

|  |  |
| --- | --- |
| **Measure Specified to Use Data From:**  **(*must be consistent with data sources entered in S.17*)** | **Measure Tested with Data From:** |
| abstracted from paper record | abstracted from paper record |
| claims | claims |
| registry | registry |
| abstracted from electronic health record | abstracted from electronic health record |
| eMeasure (HQMF) implemented in EHRs | eMeasure (HQMF) implemented in EHRs |
| other: Click here to describe | other: Click here to describe |

**1.2. If an existing dataset was used, identify the specific dataset** (*the dataset used for testing must be consistent with the measure specifications for target population and healthcare entities being measured; e.g., Medicare Part A claims, Medicaid claims, other commercial insurance, nursing home MDS, home health OASIS, clinical registry*).

STS General Thoracic Surgery Database, Version 2.2

**1.3. What are the dates of the data used in testing**? 01/01/2012 – 12/31/2014

**1.4. What levels of analysis** **were tested**? (*testing must be provided for all the levels specified and intended for measure implementation, e.g., individual clinician, hospital, health plan*)

|  |  |
| --- | --- |
| **Measure Specified to Measure Performance of:**  **(*must be consistent with levels entered in item S.20*)** | **Measure Tested at Level of:** |
| individual clinician | individual clinician |
| group/practice | group/practice |
| hospital/facility/agency | hospital/facility/agency |
| health plan | health plan |
| other: Click here to describe | other: Click here to describe |

**1.5. How many and which measured entities were included in the testing and analysis (by level of analysis and data source)**? (*identify the number and descriptive characteristics of measured entities included in the analysis (e.g., size, location, type); if a sample was used, describe how entities were selected for inclusion in the sample*)

The analysis population consisted of all STS records for patients meeting measure inclusion criteria who had their surgery during January 1, 2012 through December 31, 2014. The population included 27,844 records from 231 hospitals. Hospital-specific sample sizes ranged from 1 to 852 records per hospital (mean=121, median=85, IQR=[36, 165]).

**1.6. How many and which patients were included in the testing and analysis (by level of analysis and data source)**? (*identify the number and descriptive characteristics of patients included in the analysis (e.g., age, sex, race, diagnosis); if a sample was used, describe how patients were selected for inclusion in the sample*)

Patient Characteristics [n (%) or mean ± SD].



**1.7. If there are differences in the data or sample used for different aspects of testing (e.g., reliability, validity, exclusions, risk adjustment), identify how the data or sample are different for each aspect of testing reported below**.

The STS tests reliability based on three years of data in the General Thoracic Surgery Database (see 1.5 above). Validity testing is conducted on an annual basis through the audit of data completeness and accuracy in randomly-selected surgical records at randomly-selected GTSD participant sites (see 2b1.2 below).

**1.8** **What were the social risk factors that were available and analyzed**? For example, patient-reported data (e.g., income, education, language), proxy variables when social risk data are not collected from each patient (e.g. census tract), or patient community characteristics (e.g. percent vacant housing, crime rate) which do not have to be a proxy for patient-level data.

Patient social risk data are not collected in the General Thoracic Surgery Database. Through the collection of insurance information, information on dual Medicare/Medicaid eligibility is available from the database, which can serve as a proxy for low income and patient vulnerability. However, this information is not presently included in STS data analysis nor as a basis for stratification in STS measures.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**2a2. RELIABILITY TESTING**

***Note****: If accuracy/correctness (validity) of data elements was empirically tested*, *separate reliability testing of data elements is not required – in 2a2.1 check critical data elements; in 2a2.2 enter “see section 2b2 for validity testing of data elements”; and skip 2a2.3 and 2a2.4.*

**2a2.1. What level of reliability testing was conducted**? (*may be one or both levels*)  
 **Critical data elements used in the measure** (*e.g., inter-abstractor reliability; data element reliability must address ALL critical data elements*)  
 **Performance measure score** (e.g., *signal-to-noise analysis*)  
  
**2a2.2. For each level checked above, describe the method of reliability testing and what it tests** (*describe the steps―do not just name a method; what type of error does it test; what statistical analysis was used*)

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 estimated this quantity within the Bayesian statistical framework. We computed the squared correlation between each hospital’s estimated performance measure (the estimated SIR) and the true value (estimated using Bayesian inference methods). Accordingly, reliability was defined as the square of the Pearson correlation coefficient () between the set of participant-specific estimates

and the corresponding unknown true values, *θ*1*, . . . , θN* , that is:



The quantity was estimated by its posterior mean, namely,



where





with denoting the value of on the *l*-th MCMC sample denoting the posterior mean of .. A 95% credible interval for was obtained by calculating the 125th smallest and 125th largest values of across the 5,000 MCMC samples. 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.

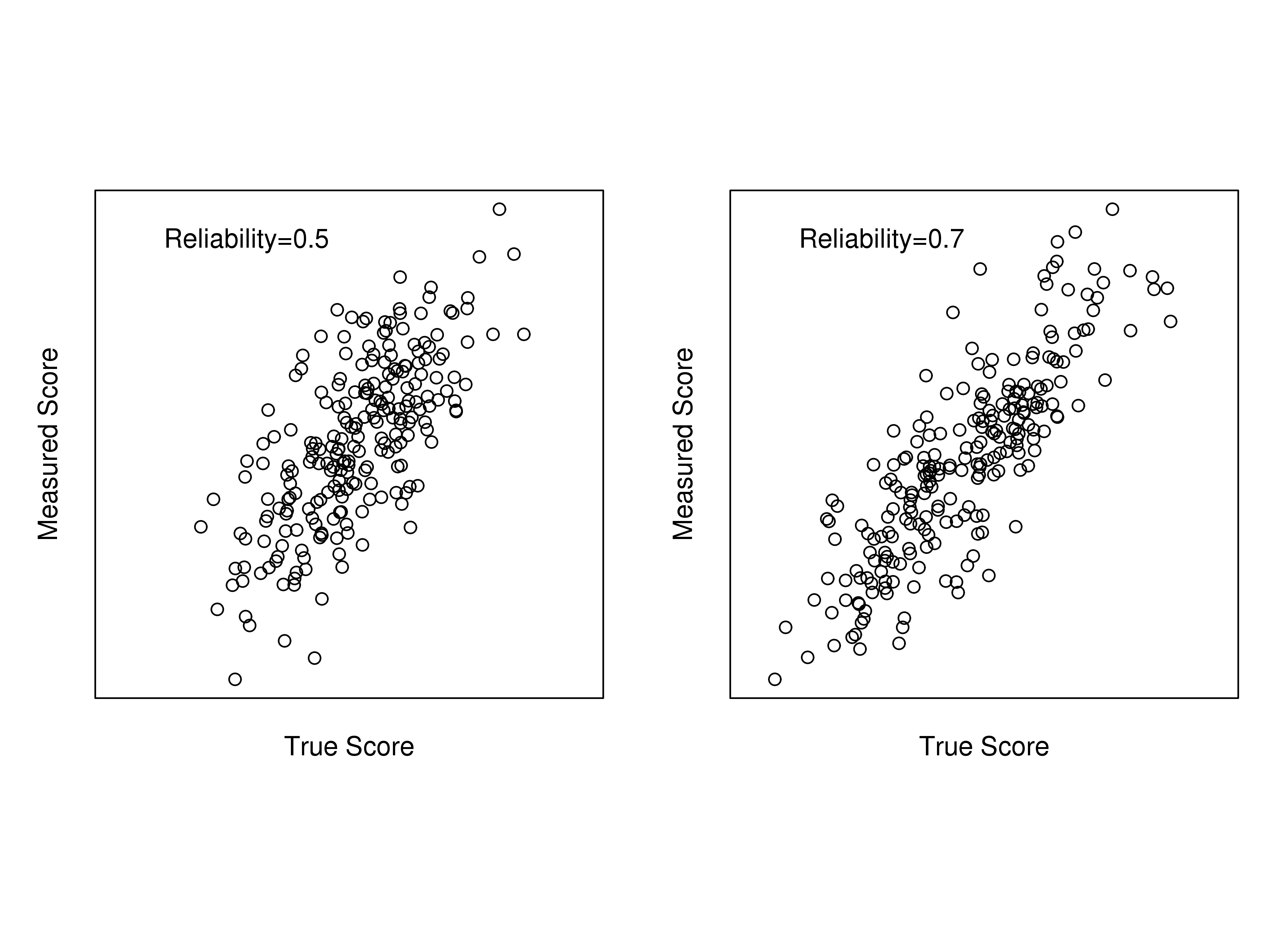
**2a2.3. For each level of testing checked above, what were the statistical results from reliability testing**? (e*.g., percent agreement and kappa for the critical data elements; distribution of reliability statistics from a signal-to-noise analysis*)  
Prior to estimating reliability, the numerical value of SIR was estimated for each hospital under the model described by Fernandez et al. (2016). The reliability measure was calculated 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 231 hospitals included, the estimate of the reliability measure is 0.50 and the 95% Bayesian probability interval (0.42, 0.58), it is 0.53 (0.45, 0.61) for 216 hospitals performing at least 10 procedures, and it is 0.84 (0.76, 0.91) for 38 hospitals with 200 or more procedures performed.

Given the timeframe of the data used for reliability testing for this measure (01/01/2012 – 12/31/2014), the revised postoperative complication data element "unexpected return to the operating room” was included in the analysis.

**2a2.4 What is your interpretation of the results in terms of demonstrating reliability**? (i*.e., what do the results mean and what are the norms for the test conducted?*)

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. Reliability increases when considering participants with increasing minimum number of cases. Starting with participants with at least 10 cases, there is a moderate reliability of 0.53, and reliability is 0.84 when only large-volume participants (at least 200 cases) are considered. The increase in reliability is the result of a more precise estimation of a participant’s measure value; in other words with the same between-participants variability, the reliability increases when the participant measurement error decreases with more cases per participant.

To visualize this effect of a decreasing measurement error on reliability, while keeping the same between-participant variability, we created two figures illustrating the accuracy of the measured scores when the true reliability is 0.50 and 0.70. Because the true score for the composite measure is unknown, we used simulated data with formula where indicates the 231 participants and where and both follow normal distributions. The standard deviations of the normal distributions were chosen such that the measure (score) has a reliability of 0.50 on the left figure and reliability of 0.70 on the right figure. Each figure has true score along the x-axis, and the estimated (measured) value of this true score along the y-axis. With a decreasing measurement error of the score (as is the case with increase in the number of cases per participant), the correlation between the true and measured values of the score increases, and thus also, equivalently, the reliability increases because reliability can be expressed as a square of this correlation (Pearson correlation). Although a high reliability of 0.70 shows a very close correlation between true and measured scores, a more moderate reliability of 0.50 still visualizes a strong association (correlation) between the true and measured values of the score.



\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**2b1. VALIDITY TESTING**

**2b1.1. What level of validity testing was conducted**? (*may be one or both levels*)  
 **Critical data elements** (*data element validity must address ALL critical data elements*)

**Performance measure score**

**Empirical validity testing** **Systematic assessment of face validity of performance measure score as an indicator** of quality or resource use (*i.e., is an accurate reflection of performance on quality or resource use and can distinguish good from poor performance*) **NOTE**: Empirical validity testing is expected at time of maintenance review; if not possible, justification is required.

**2b1.2. For each level of testing checked above, describe the method of validity testing and what it tests** (*describe the steps―do not just name a method; what was tested, e.g., accuracy of data elements compared to authoritative source, relationship to another measure as expected; what statistical analysis was used)*  
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, the STS has contracted with Telligen (formerly IFMC) and, most recently, Cardiac Registry Support, LLC (CRS) to conduct 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. Currently, auditors validate case inclusion and 15 lobectomy and 5 esophagectomy cancer cases are randomly chosen for review of 39 individual data elements. The auditors abstract each designated medical record to validate data elements previously submitted to the STS data warehouse. Agreement rates are calculated for each of the 39 elements as well as for an overall agreement rate. Five sites were randomly selected for the first audit, which took place in 2010. In 2016, 25 sites were audited.

**2b1.3. What were the statistical results from validity testing**? (*e.g., correlation; t-test*)  
STS audited 10% of participants in the General Thoracic Surgery Database in 2016 using an independent auditing firm (CRS). 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 15 randomly chosen lobectomy cancer cases and 5 esophagectomy cancer cases, comparing 39 data elements in the medical chart with the data file submitted to the STS GTSD. The agreement rate was 96.78% for overall data accuracy in 2016, with a range in agreement from 94.3% to 99.0%.

For comparison, the overall agreement rates in 2010 and 2011 were 89.9% and 94.6%, respectively (across the 33 data elements reviewed at that time). The range in agreement was from 76.5% to 95.5% in 2010, and from 88.8% to 97.5% in 2011.

Aggregate agreement rates from the 2016 audit for each of the 39 variables (data elements) and for each of the variable categories are displayed in the table below. The STS does not have access to audit results at the level of individual surgical cases; we are therefore unable to provide the kappa statistic.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| CATEGORY | FIELD\_NAME | NUM | DEN | Agreement Rate |
| PRE-OPERATIVE EVALUATION | OVERALL\_ALL\_FIELDS | 6455 | 6738 | 95.80% |
| PRE-OPERATIVE EVALUATION | Admission Date | 497 | 500 | 99.40% |
| PRE-OPERATIVE EVALUATION | Prior Cardiothoracic Surgery | 488 | 500 | 97.60% |
| PRE-OPERATIVE EVALUATION | Pre-Op Chemo-Current Malignancy | 489 | 500 | 97.80% |
| PRE-OPERATIVE EVALUATION | Pre-Op Thoracic Radiation Therapy | 489 | 500 | 97.80% |
| PRE-OPERATIVE EVALUATION | Diabetes | 413 | 423 | 97.64% |
| PRE-OPERATIVE EVALUATION | Diabetes Therapy | 68 | 82 | 82.93% |
| PRE-OPERATIVE EVALUATION | Cigarette Smoking | 489 | 500 | 97.80% |
| PRE-OPERATIVE EVALUATION | Pulmonary Function Tests Performed | 419 | 423 | 99.05% |
| PRE-OPERATIVE EVALUATION | FEV1 Predicted | 316 | 414 | 76.33% |
| PRE-OPERATIVE EVALUATION | Zubrod Score | 491 | 500 | 98.20% |
| PRE-OPERATIVE EVALUATION | Lung Cancer | 420 | 423 | 99.29% |
| PRE-OPERATIVE EVALUATION | Clinical Staging Method-Lung- EBUS | 408 | 419 | 97.37% |
| PRE-OPERATIVE EVALUATION | Clinical Staging Method-Lung-PET or PET/CT | 397 | 419 | 94.75% |
| PRE-OPERATIVE EVALUATION | Lung Cancer Tumor Size-T | 377 | 419 | 89.98% |
| PRE-OPERATIVE EVALUATION | Lung Cancer Nodes-N | 409 | 419 | 97.61% |
| PRE-OPERATIVE EVALUATION | Esophageal Cancer | 77 | 77 | 100.00% |
| PRE-OPERATIVE EVALUATION | Clinical Staging Method- Esophageal-EUS | 69 | 75 | 92.00% |
| PRE-OPERATIVE EVALUATION | Esophageal Cancer Tumor-T | 68 | 72 | 94.44% |
| PRE-OPERATIVE EVALUATION | Clinical Diagnosis of Nodal Involvement | 71 | 73 | 97.26% |
| DIAGNOSIS AND PROCEDURES | OVERALL\_ALL FIELDS | 4842 | 4978 | 97.27% |
| DIAGNOSIS AND PROCEDURES | Category of Disease-Primary | 479 | 499 | 95.99% |
| DIAGNOSIS AND PROCEDURES | Date of Surgery | 498 | 500 | 99.60% |
| DIAGNOSIS AND PROCEDURES | Procedure Start Time | 493 | 500 | 98.60% |
| DIAGNOSIS AND PROCEDURES | Procedure End Time | 482 | 500 | 96.40% |
| DIAGNOSIS AND PROCEDURES | ASA Classification | 487 | 500 | 97.40% |
| DIAGNOSIS AND PROCEDURES | Procedure | 500 | 500 | 100.00% |
| DIAGNOSIS AND PROCEDURES | Patient Disposition | 491 | 500 | 98.20% |
| DIAGNOSIS AND PROCEDURES | Pathologic Staging-Lung Cancer-T | 405 | 419 | 96.66% |
| DIAGNOSIS AND PROCEDURES | Pathologic Staging-Lung Cancer-N | 411 | 419 | 98.09% |
| DIAGNOSIS AND PROCEDURES | Lung Cancer-Number of Nodes | 385 | 419 | 91.89% |
| DIAGNOSIS AND PROCEDURES | Pathologic Staging-Esophageal Cancer-T | 69 | 74 | 93.24% |
| DIAGNOSIS AND PROCEDURES | Pathologic Staging-Esophageal Cancer-N | 73 | 74 | 98.65% |
| DIAGNOSIS AND PROCEDURES | Esophageal Cancer-Number of Nodes | 69 | 74 | 93.24% |
| POST-OPERATIVE EVENTS | OVERALL\_ALL FIELDS | 1487 | 1500 | 99.13% |
| POST-OPERATIVE EVENTS | Unexpected Return to OR | 493 | 500 | 98.60% |
| POST-OPERATIVE EVENTS | Pneumonia | 494 | 500 | 98.80% |
| POST-OPERATIVE EVENTS | Initial Vent Support >48 Hours | 500 | 500 | 100.00% |
| DISCHARGE | OVERALL\_ALL FIELDS | 1935 | 1993 | 97.09% |
| DISCHARGE | Discharge Date | 499 | 500 | 99.80% |
| DISCHARGE | Discharge Status | 490 | 500 | 98.00% |
| DISCHARGE | Readmission within 30 Days of Discharge | 484 | 493 | 98.17% |
| DISCHARGE | Status 30 Days After Surgery | 462 | 500 | 92.40% |
|  | **OVERALL\_ALL FIELDS** | **14719** | **15209** | **96.78%** |

**2b1.4. What is your interpretation of the results in terms of demonstrating validity**? (i*.e., what do the results mean and what are the norms for the test conducted?*)  
The most recent audits of the General Thoracic Surgery Database have demonstrated a high degree of data validity. Overall data accuracy rates have increased substantially since audits of the GTSD were first conducted in 2010; agreement ranges have also narrowed, indicating greater consistency in data accuracy among audited sites.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**2b2. EXCLUSIONS ANALYSIS**

**NA**  **no exclusions — *skip to section*** [***2b4***](#section2b4)

**2b2.1. Describe the method of testing exclusions and what it tests** (*describe the steps―do not just name a method; what was tested, e.g., whether exclusions affect overall performance scores; what statistical analysis was used*)

We excluded patients with missing age, sex, discharge mortality status, pathologic stage, and predicted forced expiratory volume in 1 second. In addition patients were excluded if they had an extrapleural pneumonectomy, completion pneumonectomy, carinal pneumonectomy, occult carcinoma or benign disease on final pathology, or an urgent, emergent, or palliative operation. We believe these are clinically appropriate exclusions and are necessary to make the measure a consistent performance measure for the comparison across participants. The exclusions are precisely defined and specified.

**2b2.2. What were the statistical results from testing exclusions**? (*include overall number and percentage of individuals excluded, frequency distribution of exclusions across measured entities, and impact on performance measure scores*)

There were 216 (0.7%) patients with extrapleural pneumonectomy, completion pneumonectomy, or carinal pneumonectomy; 156 (0.5%) patients with occult carcinoma or benign disease on final pathology; 3 (0.01%) with palliative operation (ASA VI); and 1510 (5.1%) non-elective status (urgent or emergent) operations, resulting in the overall exclusion of 6.3%. Impact of these exclusions on the performance measure is likely not meaningful due to a small number of cases excluded.

**2b2.3. What is your interpretation of the results in terms of demonstrating that exclusions are needed to prevent unfair distortion of performance results?** (*i.e., the value outweighs the burden of increased data collection and analysis.*  *Note:* ***If patient preference is an exclusion****, the measure must be specified so that the effect on the performance score is transparent, e.g., scores with and without exclusion*)

For the measure to consistently quantify the surgical quality of lung resection for lung cancer per its definition, it is necessary to exclude patients if they had an extrapleural pneumonectomy, completion pneumonectomy, carinal pneumonectomy, occult carcinoma or benign disease on final pathology, or an urgent, emergent, or palliative operation.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**2b3. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES**  
***If not an intermediate or health outcome, or PRO-PM, or resource use measure, skip to section*** [***2b5***](#section2b5)***.***

**2b3.1. What method of controlling for differences in case mix is used?**

**No risk adjustment or stratification**

**Statistical risk model with** Click here to enter number of factors **risk factors**

**Stratification by** Click here to enter number of categories **risk categories**

**Other,** Click here to enter description

**2b3.1.1 If using a statistical risk model, provide detailed risk model specifications, including the risk model method, risk factors, coefficients, equations, codes with descriptors, and definitions.**

Bayesian hierarchical random-effects logistic regression modeling was used to estimate hospital-specific standardized incidence ratio (SIR) and a 95% Bayesian probability interval for SIR for each of 231 hospitals. Random-effects refers to the assumption that the provider-specific parameters of interest are assumed to arise from a specified distribution defined by parameters that are also estimated in the modelling process. This analytic method is the same method used in Fernandez, et al. (2016). Risk factors in the model were: age, sex, body mass index, hypertension, steroid therapy, congestive heart failure, coronary artery disease, peripheral vascular disease, reoperation, cerebrovascular disease, diabetes mellitus, forced expiratory volume in 1 second percent of predicted, induction therapy, renal dysfunction, cigarette smoking, Zubrod score, American Society of Anesthesiologists class, approach, pathologic stage, and procedure type.

Fernandez FG, Kosinski AS, Burfeind W, Park B, DeCamp MM, Seder C, Marshall B, Magee MJ, Wright CD, Kozower BD. The Society of Thoracic Surgeons Lung Cancer Resection Risk Model: Higher Quality Data and Superior Outcomes. Ann Thorac Surg. 2016 Aug;102(2):370-7.

**2b3.2. If an outcome or resource use component measure is not risk adjusted or stratified, provide rationale and analyses to demonstrate that controlling for differences in patient characteristics (case mix) is not needed to achieve fair comparisons across measured entities**.   
n/a

**2b3.3a. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors (clinical factors or social risk factors) used in the statistical risk model or for stratification by risk** (*e.g., potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of p<0.10; correlation of x or higher; patient factors should be present at the start of care*) **Also discuss any “ordering” of risk factor inclusion**; for example, are social risk factors added after all clinical factors?

Covariates in this model were selected a priori based on a combination of literature review and expert group consensus, and as described in Fernandez, et al. (2016). All covariates were retained in the model and were not added or removed based on a statistical variable selection algorithm.

No social risk factors were used in the statistical risk model or for stratification.

Fernandez FG, Kosinski AS, Burfeind W, Park B, DeCamp MM, Seder C, Marshall B, Magee MJ, Wright CD, Kozower BD. The Society of Thoracic Surgeons Lung Cancer Resection Risk Model: Higher Quality Data and Superior Outcomes. Ann Thorac Surg. 2016 Aug;102(2):370-7.

**2b3.3b. How was the conceptual model of how social risk impacts this outcome developed? Please check all that apply:**

**Published literature**

**Internal data analysis**

**Other (please describe)**

Expert group consensus

**2b3.4a. What were the statistical results of the analyses used to select risk factors?**

Estimated odds ratios are summarized in the table below.

Fernandez FG, Kosinski AS, Burfeind W, Park B, DeCamp MM, Seder C, Marshall B, Magee MJ, Wright CD, Kozower BD. The Society of Thoracic Surgeons Lung Cancer Resection Risk Model: Higher Quality Data and Superior Outcomes. Ann Thorac Surg. 2016 Aug;102(2):370-7.

**2b3.4b. Describe the analyses and interpretation resulting in the decision to select social risk factors** *(e.g. prevalence of the factor across measured entities, empirical association with the outcome, contribution of unique variation in the outcome, assessment of between-unit effects and within-unit effects.)* **Also describe the impact of adjusting for social risk (or not) on providers at high or low extremes of risk.**

All covariates were retained in the model and were not added or removed based on a statistical variable selection algorithm.

As noted in 1.8 above, patient social risk data are not collected in the General Thoracic Surgery Database.

**2b3.5. Describe the method of testing/analysis used to develop and validate the adequacy of the statistical model or stratification approach** (*describe the steps―do not just name a method; what statistical analysis was used*)

Continuous variables were evaluated with respect to linearity of effect and no departure from linearity was noted. The calibration of the model was assessed with the Hosmer-Lemeshow goodness-of-fit statistic. The discrimination of the model was assessed with the C-statistic.

*Provide the statistical results from testing the approach to controlling for differences in patient characteristics (case mix) below*.

***If stratified, skip to*** [***2b3.9***](#question2b49)

**2b3.6. Statistical Risk Model Discrimination Statistics** (*e.g., c-statistic, R-squared*)**:**

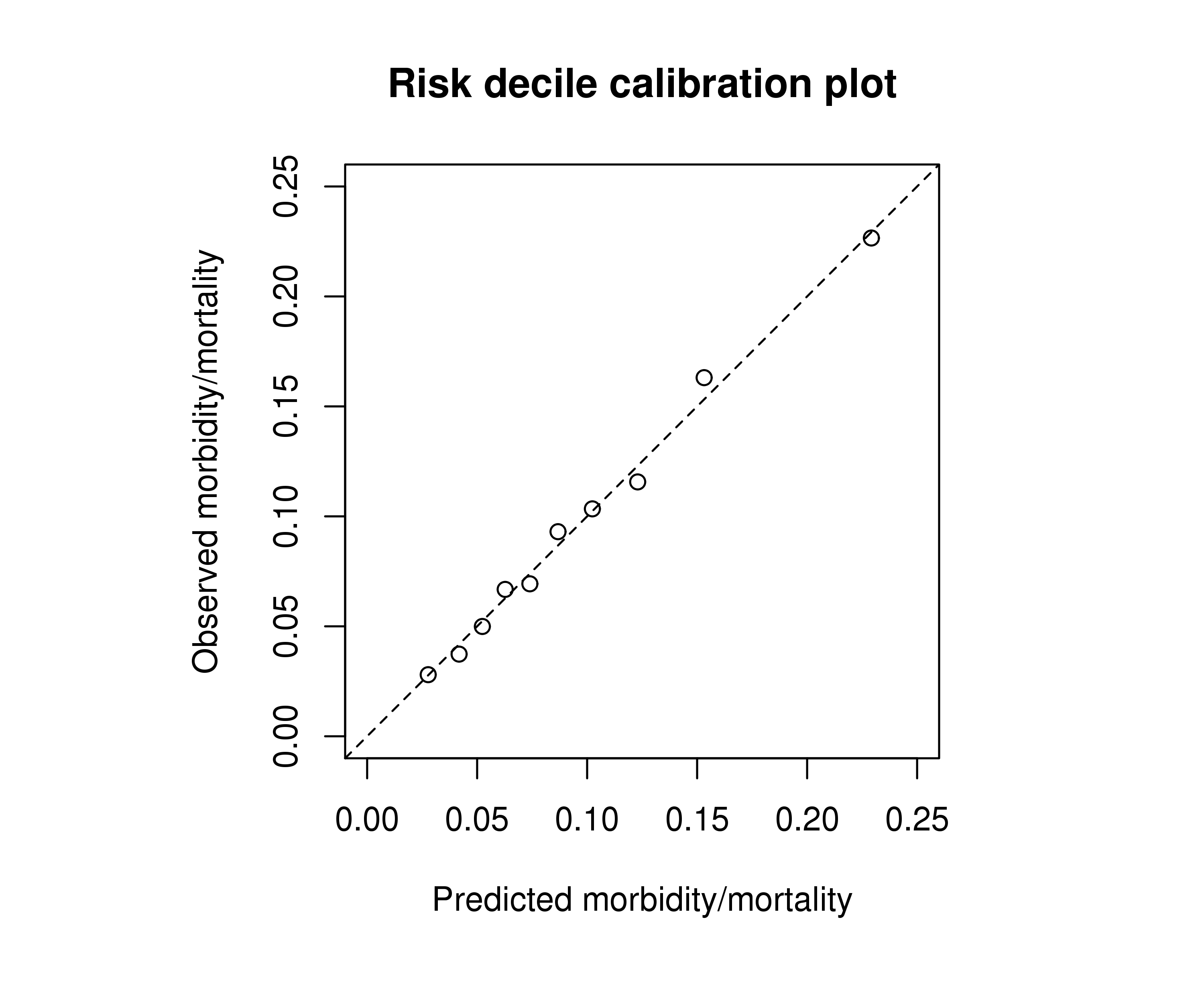
The C-statistics is 0.68.

**2b3.7. Statistical Risk Model Calibration Statistics** (*e.g., Hosmer-Lemeshow statistic*):

The Hosmer-Lemeshow goodness-of-fit p-value=0.40 demonstrates that the model estimates fit the data at an acceptable level.

**2b3.8. Statistical Risk Model Calibration – Risk decile plots or calibration curves**:

Risk decile plot below shows good alignment of predicted and observed probabilities of outcome (operative mortality or major morbidity) within deciles of predicted values.

****

**2b3.9. Results of Risk Stratification Analysis**:

n/a

**2b3.10. What is your interpretation of the results in terms of demonstrating adequacy of controlling for differences in patient characteristics (case mix)?** (i*.e., what do the results mean and what are the norms for the test conducted*)

The results demonstrated that the STS lung resection for lung cancer risk model is well calibrated and has good discrimination power. It is suitable for controlling for differences in case-mix between centers.

**2b3.11.** **Optional Additional Testing for Risk Adjustment** (*not required, but would provide additional support of adequacy of risk model, e.g., testing of risk model in another data set; sensitivity analysis for missing data; other methods that were assessed*)

n/a

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**2b4. IDENTIFICATION OF STATISTICALLY SIGNIFICANT & MEANINGFUL DIFFERENCES IN PERFORMANCE**

**2b4.1. Describe the method for determining if statistically significant and clinically/practically meaningful differences in performance measure scores among the measured entities can be identified** (*describe the steps―do not just name a method; what statistical analysis was used? Do not just repeat the information provided related to performance gap in 1b)*

Bayesian hierarchical modeling was used to estimate hospital-specific standardized incidence ratio (SIR) and a 95% Bayesian probability interval for SIR for each of 231 hospitals. The degree of uncertainty surrounding an STS participant’s SIR is indicated by calculating 95% Bayesian credible intervals (CrI’s) which are similar to conventional confidence intervals. An STS participant’s performance is considered average if the Bayesian credible interval (CrI) surrounding their SIR score overlaps 1. If the Bayesian CrI falls entirely below 1, the participant has lower-than-expected performance. If the Bayesian CrI falls entirely above 1, the participant has higher-than-expected performance.

**2b4.2. What were the statistical results from testing the ability to identify statistically significant and/or clinically/practically meaningful differences in performance measure scores across measured entities?** (e.g., *number and percentage of entities with scores that were statistically significantly different from mean or some benchmark, different from expected; how was meaningful difference defined*)

Figure 1 under the Results section of the attachment (Fernandez et al, 2016) displays estimated SIR and corresponding 95% Bayesian probability interval for each of 231 hospitals. Hospitals are ordered according to the increasing SIR estimate. There are meaningful differences between the best performing (3.5%; 8 of 231 sites) and the worst performing hospitals (6.9%; 16 of 231 sites). This indicates that this model provides meaningful discrimination between best and worst performers.

**2b4.3. What is your interpretation of the results in terms of demonstrating the ability to identify statistically significant and/or clinically/practically meaningful differences in performance across measured entities?** (i*.e., what do the results mean in terms of statistical and meaningful differences?*)

The identified differences in performance between centers are both statistically significant and clinically meaningful. The surgeon panel and users are satisfied with the distribution of participants across performance categories.

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**2b5. COMPARABILITY OF PERFORMANCE SCORES WHEN MORE THAN ONE SET OF SPECIFICATIONS**

***If only one set of specifications, this section can be skipped.***

**Note***: This item is directed to measures that are risk-adjusted (with or without social risk factors)* ***OR*** *to measures with more than one set of specifications/instructions (e.g., one set of specifications for how to identify and compute the measure from medical record abstraction and a different set of specifications for claims or eMeasures). It does not apply to measures that use more than one source of data in one set of specifications/instructions (e.g., claims data to identify the denominator and medical record abstraction for the numerator).* ***Comparability is not required when comparing performance scores with and without social risk factors in the risk adjustment model. However, if comparability is not demonstrated for measures with more than one set of specifications/instructions, the different specifications (e.g., for medical records vs. claims) should be submitted as separate measures.***

**2b5.1. Describe the method of testing conducted to compare performance scores for the same entities across the different data sources/specifications** (*describe the steps―do not just name a method; what statistical analysis was used*)  
n/a

**2b5.2. What were the statistical results from testing comparability of performance scores for the same entities when using different data sources/specifications?** (*e.g., correlation, rank order*)  
n/a

**2b5.3. What is your interpretation of the results in terms of the differences in performance measure scores for the same entities across the different data sources/specifications?** (i*.e., what do the results mean and what are the norms for the test conducted*)  
n/a

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**2b6. MISSING DATA ANALYSIS AND MINIMIZING BIAS**

**2b6.1. Describe the method of testing conducted to identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased** due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias (*describe the steps―do not just name a method; what statistical analysis was used*)

The quality of data in STS General Thoracic Surgery Database has been improving. We managed the remaining missing data with imputation. Missing body mass index (BMI) values (1%) were imputed utilizing sex specific median of the observed BMI values. For binary risk factors, missing values were considered as indicating absence of the risk factor.

**2b6.2. What is the overall frequency of missing data, the distribution of missing data across providers, and the results from testing related to missing data?** (*e.g.,**results of sensitivity analysis of the effect of various rules for missing data/nonresponse; if no empirical sensitivity analysis, identify the approaches for handling missing data that were considered and pros and cons of each*)

Patients with missing age, sex, discharge mortality status, pathologic stage, and predicted forced expiratory volume in 1 second were excluded. All the variables in the population utilized for this measure had less than 1% of missing values.

**2b6.3. What is your interpretation of the results in terms of demonstrating that performance results are not biased** due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias**?** (i*.e., what do the results mean in terms of supporting the selected approach for missing data and what are the norms for the test conducted; if no empirical analysis, provide rationale for the selected approach for missing data*)

The rates of missing data were low. We therefore concluded that systematic missing data did not lead to bias in our measure.