NATIONAL QUALITY FORUM

MEASURE WORKSHEET

This document summarizes the evaluation of the measure as it progresses through NQF's Consensus Development Process (CDP). The information submitted by measure developers/stewards is included after the Brief Measure Information, Preliminary Analysis, and Pre-meeting Public and Member Comments sections.

To navigate the links in the worksheet: Click to go to the link. ALT + LEFT ARROW to return

Purple text represents the responses from measure developers.

Red text denotes developer information that has changed since the last measure evaluation review.

Brief Measure Information

NQF #: 1790

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

Measure Steward: The Society of Thoracic Surgeons

Brief Description of Measure: Percentage of patients greater than or equal to 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, unexpected return to the operating room, myocardial infarction or operative mortality (death during the index hospitalization, regardless of timing, or within 30 days, regardless of location).

Developer Rationale: 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 enable them to compare risk-adjusted outcomes for quality improvement purposes.

Numerator Statement: Number of patients greater than or equal to 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, unexpected return to the operating room, myocardial infarction or operative mortality (death during the index hospitalization, regardless of timing, or within 30 days, regardless of location).

Denominator Statement: Number of patients greater than or equal to 18 years of age undergoing elective lung resection (Open or VATS wedge resection, segmentectomy, lobectomy, bilobectomy, sleeve lobectomy, pneumonectomy) for lung cancer

Denominator Exclusions: 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. Furthermore, patients with missing age, sex, discharge mortality status, and predicted forced expiratory volume in 1 second were also excluded.

Measure Type: Outcome

Data Source: Other, Registry Data

Level of Analysis: Facility

IF Endorsement Maintenance – Original Endorsement Date: Aug 09, 2012 **Most Recent Endorsement Date:** Aug 09, 2012

Staff Preliminary Analysis: Maintenance of Endorsement

To maintain NQF endorsement endorsed measures are evaluated periodically to ensure that the measures still meets the NQF endorsement criteria ("maintenance"). The emphasis for maintaining endorsement is focused on how effective the measure is for promoting improvements in quality. Endorsed measures should have some experience from the field to inform the evaluation. The emphasis for maintaining endorsement is noted for each criterion.

Criteria 1: Importance to Measure and Report

1a. Evidence

Maintenance measures – less emphasis on evidence unless there is new information or change in evidence since the prior evaluation.

1a. Evidence. The evidence requirements for a health outcome measure include providing empirical data that demonstrate a relationship between the outcome and at least one healthcare structure, process, intervention, or service; if these data not available, data demonstrating wide variation in performance, assuming the data are from a robust number of providers and results are not subject to systematic bias. For measures derived from patient report, evidence also should demonstrate that the target population values the measured outcome, process, or structure and finds it meaningful.

Summary of prior review in 2012

- This measure assesses postoperative complications and operative mortality during lung cancer resection surgery. In the prior review, the Committee agreed that the evidence was solid and demonstrated substantial variation in morbidity and mortality after lung cancer surgery.

Changes to evidence from last review

☐ **The developer attests that there have been no changes in the evidence since the measure was last evaluated.**

☒ **The developer provided updated evidence for this measure:**

Updates:

- The developer provided updated evidence (Fernandez et al., 2016) on the STS lung cancer resection risk model which identifies predictors of complications and mortality including patient age, smoking status, comorbid medical conditions, and other patient characteristics. Fernandez et al concluded that operative mortality and complication rates are low for lung cancer resection among surgeons participating in the STS General Thoracic Surgery Database. The developer reports that "knowledge of these predictors informs clinical decision making by enabling physicians and patients to understand the association between patient characteristics and outcomes".
- The developer provided performance data for 217,844 patient records at 213 sites from January 1, 2012 through December 31, 2014 demonstrating a variation in performance from 0.47% to 2.37%.
- *Empirical data* demonstrating a relationship between the outcome to at least one healthcare process is now required. NQF guidance states that a wide variation in performance can be used as evidence, assuming the data are from a robust number of providers and results are not subject to systematic bias.

Question for the Committee:

- *Is there at least one thing that the provider can do to achieve a change in the measure results?*
- *Is the performance data sufficient, in size and variance, to demonstrate that some hospitals are engaging in quality improvement activities to decrease morbidity and mortality in lung cancer patients undergoing elective lung resection better than others?*

Guidance from the Evidence Algorithm: Measure assesses performance on a health outcome (Box 1) → There is a relationship between the health outcome and one healthcare action (Box 2) → Pass

Preliminary rating for evidence: ☒ **Pass** ☐ **No Pass**

1b. Gap in Care/Opportunity for Improvement and 1b. Disparities

Maintenance measures – increased emphasis on gap and variation

1b. Performance Gap. The performance gap requirements include demonstrating quality problems and opportunity for improvement.

- The developer reports that there is no overlap for the hospital specific standardized incidence ratio (SIR) between the best performing sites (3.5%; 8 of 231 sites with upper limit *below* 1) and worst performing sites (6.9%: 16 of 231 sites with lower limit *above* 1). SIR were calculated for 27,844 patient records at 213 sites during January 1, 2012 through December 31, 2014. The distribution of hospital specific estimates of the SIR for morbidity and mortality is shown below.

Minimum	**0.47**
1st quartile	0.90
Median	1.00
Mean	1.05
3rd quartile	1.22
Maximum	2.37

Disparities

- Using the same data described above, incidence of mortality or major morbidity was calculated for race:

Race, N	**%**	**Confidence interval**
White, N=24,099	9.8	95% [9.4, 10.1]
Black, N=2,369	8.9	95% [7.8,10.1]
Other, N=1,217	6.9	95% [5.6, 8.5]

Questions for the Committee:

- *Does the measure demonstrate a quality problem related to morbidity and mortality in lung cancer patients undergoing elective lung resection?*
- *Is a national performance still warranted?*
- *Are you aware of evidence that other disparities exist in this area of healthcare?*

Preliminary rating for opportunity for improvement: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee Pre-evaluation Comments: Criteria 1: Importance to Measure and Report (including 1a, 1b, 1c)

1a. Evidence to Support Measure Focus

**Ample evidence from the STS registry is provided to calculate this measure.

** Evidence well supported

** Outcome measure with good data to support it

** This maintenance measure uses the well established Society for Thoracic Surgery risk adjusted database to evaluate the mortality and major morbidities after lung resection for lung cancer. Studies are cited that address the data integrity and utility of the measure.

1b. Performance Gap

**Gap is relatively small (3.5% good performers and 6.9% bad performers) but this measure is the most important outcome of surgery and therefore important for public accountability.

**Performance gap present

**There remains a performance gap for this existing measure

****A performance gap was demonstrated by the almost 5 fold difference between high-performing and low performing hospitals. Only racial disparities were addressed in the submission.

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability: Specifications **and** Reliability
2b. Validity: Testing; Exclusions**;** Risk-Adjustment**;** Meaningful Differences**;** Comparability Missing Data

Reliability

2a1. Specifications requires the measure, as specified, to produce consistent (reliable) and credible (valid) results about the quality of care when implemented. For maintenance measures – no change in emphasis – specifications should be evaluated the same as with new measures.

2a2. Reliability testing demonstrates if 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 enough to distinguish differences in performance across providers. For maintenance measures – less emphasis if no new testing data provided.

Validity

2b2. Validity testing should demonstrate the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality. For maintenance measures – less emphasis if no new testing data provided.

2b2-2b6. Potential threats to validity should be assessed/addressed.

Composite measures only:

2d. Empirical analysis to support composite construction. Empirical analysis should demonstrate that the component measures add value to the composite and that the aggregation and weighting rules are consistent with the quality construct.

Complex measure evaluated by Scientific Methods Panel? ☒ **Yes** ☐ **No**

Evaluators: Michael Stoto, Zhenqiu Lin, Susan White

Evaluation of Reliability and Validity (and composite construction, if applicable):

Evaluation A

Evaluation B

Evaluation C

Questions for the Committee regarding reliability:

- *Do you have any concerns that the measure can be consistently implemented (i.e., are measure specifications adequate)?*
- *The Scientific Methods Panel is satisfied with the reliability testing for the measure. Does the Committee think there is a need to discuss and/or vote on reliability?*

Questions for the Committee regarding validity:

- *Do you have any concerns regarding the validity of the measure (e.g., exclusions, risk-adjustment approach, etc.)?*

o *The Scientific Methods Panel is satisfied with the validity analyses for the measure. Does the Committee think there is a need to discuss and/or vote on validity?*

Preliminary rating for reliability:	☐ **High**	☒ **Moderate**	☐ **Low**	☐ **Insufficient**
Preliminary rating for validity:	☐ **High**	☒ **Moderate**	☐ **Low**	☐ **Insufficient**

Evaluation A: Scientific Acceptability

Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. **Measures must be judged to meet the subcriteria for both reliability and validity to pass this criterion.**

Instructions:

- Please complete this form for each measure you are evaluating.
- Please pay close attention to the skip logic directions.
- If you are unable to check a box, please highlight or shade the box for your response.
- You must answer the "overall rating" item for both Reliability and Validity. Also, be sure to answer the composite measure question at the end of the form if your measure is a composite.
- We have provided TIPS to help you answer the questions.
- We've designed this form to try to minimize the amount of writing that you have to do. That said, ***it is critical that you explain your thinking/rationale if you check boxes where we ask for an explanation*** (because this is a Word document, you can just add your explanation below the checkbox)**.** Feel free to add additional explanation, even if an explanation is not requested (but please type this underneath the appropriate checkbox).
- This form is based on Algorithms 2 and 3 in the Measure Evaluation Criteria and Guidance document (see pages 18-24). These algorithms provide guidance to help you rate the Reliability and Validity subcriteria. ***We ask that you refer to this document when you are evaluating your measures***.
- Please contact Methods Panel staff if you have questions (methodspanel@qualityforum.org).

Measure Number: 1790

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

RELIABILITY

1. Are submitted specifications precise, unambiguous, and complete so that they can be consistently implemented?
   NOTE: NQF staff will conduct a separate, more technical, check of eMeasure (eCQM) specifications, value sets, logic, and feasibility, so no need to consider these in your evaluation.

   TIPS: Consider the following: Are all the data elements clearly defined? Are all appropriate codes included? Is the logic or calculation algorithm clear? Is it likely this measure can be consistently implemented?

   ☒**Yes (go to Question #2)**

   ☐No (please explain below, and go to Question #2) NOTE that even though ***non-precise specifications should result in an overall LOW rating for reliability***, we still want you to look at the testing results.

2. Was empirical reliability testing (at the data element or measure score level) conducted using statistical tests with the measure as specified?

   TIPS: Check the 2nd "NO" box below if: only descriptive statistics provided; only describes process for data management/cleaning/computer programming; testing does not match measure specifications (i.e. data, eMeasure, level of analysis, patients)

   ☒**Yes (go to Question #4)**

   ☐No, there is reliability testing information, but *not* using statistical tests and/or not for the measure as specified OR there is no reliability testing (please explain below then go to

Question #3)

3. Was **empirical VALIDITY testing** of patient-level data conducted?

   ☐Yes (use your rating from data element validity testing – Question #16- under Validity Section)

   ☐No (please explain below and rate Question #11: OVERALL RELIABILITY as INSUFFICIENT and proceed to the VALIDITY SECTION)

4. Was reliability testing conducted with computed performance measure scores for each measured entity?

   TIPS: Answer no if: only one overall score for all patients in sample used for testing patient-level data

   ☒**Yes (go to Question #5)**

   ☐No (go to Question #8)

5. Was the method described and appropriate for assessing the proportion of variability due to real differences among measured entities? *NOTE: If multiple methods used, at least one must be appropriate.*

   TIPS: Examples of appropriate methods include signal-to-noise analysis (e.g. Adams/RAND tutorial); random split-half correlation; other accepted method with description of how it assesses reliability of the performance score.

   ☒**Yes (go to Question #6)**

   ☐No (please explain below then go to Question #8)

6. **RATING (score level)** - What is the level of certainty or confidence that the performance measure scores are reliable?

   TIPS: Consider the following: Is the test sample adequate to generalize for widespread implementation? Do the results demonstrate sufficient reliability so that differences in performance can be identified?

   ☒**High (go to Question #8)**

   ☐Moderate (go to Question #8)

   ☐Low (please explain below then go to Question #7)

7. Was other reliability testing reported?

   ☐Yes (go to Question #8)

   ☐No (rate Question #11: OVERALL RELIABILITY as LOW and proceed to the VALIDITY SECTION)

8. Was reliability testing conducted with patient-level data elements that are used to construct the performance measure?

   TIPS: Prior reliability studies of the same data elements may be submitted; if comparing abstraction to "authoritative source/gold standard" see Validity Section Question #15)

   ☐Yes (go to Question #9)

   ☒**No (if there is score-level testing, rate Question #11: OVERALL RELIABILITY based on score-level rating from Question #6; otherwise, rate Question #11: OVERALL RELIABILITY as INSUFFICIENT. Then proceed to the** VALIDITY SECTION**)**

9. Was the method described and appropriate for assessing the reliability of ALL critical data elements?

   TIPS: For example: inter-abstractor agreement (ICC, Kappa); other accepted method with description of how it assesses reliability of the data elements

   Answer no if: only assessed percent agreement; did not assess separately for all data elements (at least numerator, denominator, exclusions)

   ☐Yes (go to Question #10)

   ☐No (if no, please explain below and rate Question #10 as INSUFFICIENT)

10. **RATING (data element)** – Based on the reliability statistic and scope of testing (number and representativeness of patients and entities), what is the level of certainty or confidence that the data used in the measure are reliable?

    TIPS: Consider the following: Is the test sample adequate to generalize for widespread implementation? Can data elements be collected consistently?

    ☐Moderate (if score-level testing was NOT conducted, rate Question #11: OVERALL RELIABILITY as MODERATE)

    ☐Low (if score-level testing was NOT conducted, rate Question #11: OVERALL RELIABILITY as LOW)

    ☐Insufficient (go to Question #11)

11. OVERALL RELIABILITY RATING

OVERALL RATING OF RELIABILITY taking into account precision of specifications and all testing results:

☒**High (NOTE: Can be HIGH only if score-level testing has been conducted)**

☐Moderate (NOTE: Moderate is the highest eligible rating if score-level testing has not been conducted)

☐Low (please explain below) [NOTE: Should rate LOW if you believe specifications are NOT precise, unambiguous, and complete]

☐Insufficient (please explain below) [NOTE: For most measure types, testing at both the score level and the data element level is not required]

Please note that reliability testing was conducted at the level of the hospital, but not at the clinician or group/practice level, so my conclusions apply only to the hospital level results.

VALIDITY

ASSESSMENT OF THREATS TO VALIDITY

1. Were all potential threats to validity that are relevant to the measure empirically assessed?

   TIPS: Threats to validity include: exclusions; need for risk adjustment; Able to identify statistically significant and meaningful differences; multiple sets of specifications; missing data/nonresponse.

   ☒**Yes (go to Question #2)**

   ☐No (please explain below and go to Question #2) [NOTE that even if ***non-assessment of applicable threats should result in an overall INSUFFICENT rating for validity***, we still want you to look at the testing results]

2. Analysis of potential threats to validity: Any concerns with measure exclusions?

   TIPS: Consider the following: Are the exclusions consistent with the evidence? Are any patients or patient groups inappropriately excluded from the measure? Are the exclusions/exceptions of sufficient frequency and variation across providers to be needed (and outweigh the data collection burden)? If patient preference (e.g., informed decisionmaking) is a basis for exclusion, does it impact performance and if yes, is the measure specified so that the information about patient preference and the effect on the measure is transparent?

   ☐Yes (please explain below then go to Question #3)

   ☒**No (go to Question #3)**

   ☐Not applicable (i.e., there are no exclusions specified for the measure; go to Question #3)

3. Analysis of potential threats to validity: Risk-adjustment (applies to all outcome, cost, and resource use measures; may also apply to other types of measure)

   ☐Not applicable (e.g., structure or process measure that is not risk-adjusted; go to Question #4)

   a. Is a conceptual rationale for social risk factors included? ☐Yes ☒**No**

b. Are social risk factors included in risk model? ☐Yes ☒**No**

c. Any concerns regarding the risk-adjustment approach?

*TIPS: Consider the following: If a justification for **not risk adjusting** is provided, is there any evidence that contradicts the developer's rationale and analysis? If the developer asserts there is **no conceptual basis** for adjusting this measure for social risk factors, do you agree with the rationale? **If risk adjusted**: Are the candidate and final variables included in the risk adjustment model adequately described for the measure to be implemented? Are the candidate and final variables included in the risk adjustment model adequately described for the measure to be implemented? Are all of the risk adjustment variables present at the start of care (if not, do you agree with the rationale)? If social risk factors are not included in the risk-adjustment approach, do you agree with the developer's decision? Is an appropriate risk-adjustment strategy included in the measure (e.g., adequate model discrimination and calibration)? Are all statistical model specifications included, including a "clinical model only" if social risk factors are included in the final model?*

☒**Yes (please explain below then go to Question #4)**

☐No (go to Question #4)

Adjustment for social risk factors was not done or even discussed.

4. Analysis of potential threats to validity: Any concerns regarding ability to identify meaningful differences in performance or overall poor performance?

   ☐Yes (please explain below then go to Question #5)

   ☒**No (go to Question #5)**

5. Analysis of potential threats to validity: Any concerns regarding comparability of results if multiple data sources or methods are specified?

   ☐Yes (please explain below then go to Question #6)

   ☒**No (go to Question #6)**

   ☐Not applicable (go to Question #6)

6. Analysis of potential threats to validity: Any concerns regarding missing data?

   ☐Yes (please explain below then go to Question #7)

   ☒**No (go to Question #7)**

ASSESSMENT OF MEASURE TESTING

7. Was <u>empirical</u> validity testing conducted using the measure as specified and appropriate statistical test?

   Answer no if: face validity; only refer to clinical evidence; only descriptive statistics; only describe process for data management/cleaning/computer programming; testing does not match measure specifications (i.e. data, eMeasure, level, setting, patients).

   ☒**Yes (go to Question #10**) [NOTE: If appropriate empirical testing has been conducted, then evaluation of face validity is not necessary. Go to Question #8 **only if** there is insufficient information provided to evaluate data element and score-level testing.]

   ☐No (please explain below then go to Question #8)

8. Was <u>face validity</u> systematically assessed by recognized experts to determine agreement on whether the computed performance measure score from the measure as specified can be used to distinguish good and poor quality?

   TIPS: Answer no if: focused on data element accuracy/availability/feasibility/other topics; the degree of consensus and any areas of disagreement not provided/discussed.

   ☐Yes (go to Question #9)

☐No (please explain below and rate Question #17: OVERALL VALIDITY as INSUFFICIENT)

9. **RATING (face validity)** - Do the face validity testing results indicate substantial agreement that the performance measure score from the measure as specified can be used to distinguish quality AND potential threats to validity are not a problem, OR are adequately addressed so results are not biased?

   ☐Yes (if a NEW measure, rate Question #17: OVERALL VALIDITY as MODERATE)

   ☐ Yes (if a MAINTENANCE measure, do you agree with the justification for not conducting empirical testing? If no, rate Question #17: OVERALL VALIDITY as INSUFFICIENT; otherwise, rate Question #17: OVERALL VALIDITY as MODERATE)

   ☐No (please explain below and rate Question #17: OVERALL VALIDITY AS LOW)

10. Was validity testing conducted with computed performance measure scores for each measured entity?

    TIPS: Answer no if: one overall score for all patients in sample used for testing patient-level data.

    ☒**Yes (go to Question #11)**

    ☐No (please explain below and go to Question #13)

Please note that reliability testing was conducted at the level of the hospital, but not at the clinician or group/practice level, so my conclusions apply only to the hospital level results.

11. Was the method described and appropriate for assessing conceptually and theoretically sound hypothesized relationships?

    TIPS: For example: correlation of the performance measure score on this measure and other performance measures; differences in performance scores between groups known to differ on quality; other accepted method with description of how it assesses validity of the performance score

    ☒**Yes (go to Question #12)**

    ☐No (please explain below, rate Question #12 as INSUFFICIENT and then go to Question #14)

12. **RATING (measure score)** - Based on the measure score results (significance, strength) and scope of testing (number of measured entities and representativeness) and analysis of potential threats, what is the level of certainty or confidence that the performance measure scores are a valid indicator of quality?

    ☐High (go to Question #14)

    ☒**Moderate (go to Question #14)**

    ☐Low (please explain below then go to Question #13)

    ☐Insufficient

13. Was other validity testing reported?

    ☒**Yes (go to Question #14)**

    ☐No (please explain below and rate Question #17: OVERALL VALIDITY as LOW)

14. Was validity testing conducted with patient-level data elements?

    TIPS: Prior validity studies of the same data elements may be submitted

    ☒**Yes (go to Question #15)**

    ☐No (please explain below and rate Question #17: OVERALL VALIDITY as INSUFFICIENT if no score-level testing was conducted, otherwise, rate Question #17: OVERALL VALIDITY based on score-level rating from Question #12)

15. Was the method described and appropriate for assessing the accuracy of ALL critical data elements? *NOTE that data element validation from the literature is acceptable.*

TIPS: For example: Data validity/accuracy as compared to authoritative source- sensitivity, specificity, PPV, NPV; other accepted method with description of how it assesses validity of the data elements.

Answer No if: only assessed percent agreement; did not assess separately for all data elements (at least numerator, denominator, exclusions)

☒**Yes (go to Question #16)**

☐No (please explain below and rate Question #16 as INSUFFICIENT)

The developers reported only percent agreement rather than sensitivity/specificity and positive/negative predictive values. However, since the percent agreement figures were so consistently high (96.78% overall with a range from 94.3% to 99.0%), I believe that the analysis is sufficient to rate the data element validity as Moderate.

16. **RATING (data element)** - Based on the data element testing results (significance, strength) and scope of testing (number and representativeness of patients and entities) and analysis of potential threats, what is the level of certainty or confidence that the data used in the measure are valid?

☒**Moderate (if score-level testing was NOT conducted, rate Question #17: OVERALL VALIDITY as MODERATE)**

☐Low (please explain below) (if score-level testing was NOT conducted, rate Question #17: OVERALL VALIDITY as LOW)

☐Insufficient (go to Question #17)

17. OVERALL VALIDITY RATING

OVERALL RATING OF VALIDITY taking into account the results and scope of all testing and analysis of potential threats.

☐High (NOTE: Can be HIGH only if score-level testing has been conducted)

☒**Moderate (NOTE: Moderate is the highest eligible rating if score-level testing has NOT been conducted)**

☒Low (please explain below) [NOTE: Should rate LOW if you believe that there are threats to validity and/or threats to validity were not assessed]

☐Insufficient (if insufficient, please explain below) [NOTE: For most measure types, testing at both the score level and the data element level is not required] [NOTE: If rating is INSUFFICIENT for all empirical testing, then go back to Question #8 and evaluate any face validity that was conducted, then reconsider this overall rating.]

Please note that reliability testing was conducted at the level of the hospital, but not at the clinician or group/practice level, so my conclusions apply only to the hospital level results.

This is rated Moderate rather than High because there is no adjustment for social factors.

FOR COMPOSITE MEASURES ONLY: Empirical analyses to support composite construction

What is the level of certainty or confidence that the empirical analysis demonstrates that the component measures add value to the composite and that the aggregation and weighting rules are consistent with the quality construct?

TIPS: Consider the following: Do the component measures fit the quality construct? Are the objectives of parsimony and simplicity achieved while supporting the quality construct?

☐High

☐Moderate

☐Low (please explain below)

☐Insufficient (please explain below)

Evaluation B: Scientific Acceptability

Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. **Measures must be judged to meet the subcriteria for both reliability and validity to pass this criterion.**

Instructions:

- Please complete this form for each measure you are evaluating.
- Please pay close attention to the skip logic directions.
- If you are unable to check a box, please highlight or shade the box for your response.
- You must answer the "overall rating" item for both Reliability and Validity. Also, be sure to answer the composite measure question at the end of the form if your measure is a composite.
- We have provided TIPS to help you answer the questions.
- We've designed this form to try to minimize the amount of writing that you have to do. That said, ***it is critical that you explain your thinking/rationale if you check boxes where we ask for an explanation*** (because this is a Word document, you can just add your explanation below the checkbox)**.** Feel free to add additional explanation, even if an explanation is not requested (but please type this underneath the appropriate checkbox).
- This form is based on Algorithms 2 and 3 in the Measure Evaluation Criteria and Guidance document (see pages 18-24). These algorithms provide guidance to help you rate the Reliability and Validity subcriteria. ***We ask that you refer to this document when you are evaluating your measures***.
- Please contact Methods Panel staff if you have questions (methodspanel@qualityforum.org).

Measure Number: 1790

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

RELIABILITY

11. Are submitted specifications precise, unambiguous, and complete so that they can be consistently implemented?

    NOTE: NQF staff will conduct a separate, more technical, check of eMeasure (eCQM) specifications, value sets, logic, and feasibility, so no need to consider these in your evaluation.

    TIPS: Consider the following: Are all the data elements clearly defined? Are all appropriate codes included? Is the logic or calculation algorithm clear? Is it likely this measure can be consistently implemented?

    ☒Yes (go to Question #2)

    ☐No (please explain below, and go to Question #2) NOTE that even though ***non-precise specifications should result in an overall LOW rating for reliability***, we still want you to look at the testing results.

12. Was empirical reliability testing (at the data element or measure score level) conducted using statistical tests with the measure as specified?

    TIPS: Check the 2nd "NO" box below if: only descriptive statistics provided; only describes process for data management/cleaning/computer programming; testing does not match measure specifications (i.e. data, eMeasure, level of analysis, patients)

    ☒Yes (go to Question #4)

    ☐No, there is reliability testing information, but *not* using statistical tests and/or not for the measure as specified OR there is no reliability testing (please explain below then go to Question #3)

13. Was **empirical VALIDITY testing** of patient-level data conducted?

    ☐Yes (use your rating from data element validity testing – Question #16- under Validity Section)

☐No (please explain below and rate Question #11: OVERALL RELIABILITY as INSUFFICIENT and proceed to the VALIDITY SECTION)

14. Was reliability testing conducted with computed performance measure scores for each measured entity?

    TIPS: Answer no if: only one overall score for all patients in sample used for testing patient-level data

    ☒Yes (go to Question #5)

    ☐No (go to Question #8)

15. Was the method described and appropriate for assessing the proportion of variability due to real differences among measured entities? *NOTE: If multiple methods used, at least one must be appropriate.*

    TIPS: Examples of appropriate methods include signal-to-noise analysis (e.g. Adams/RAND tutorial); random split-half correlation; other accepted method with description of how it assesses reliability of the performance score.

    ☒Yes (go to Question #6)

    ☐No (please explain below then go to Question #8)

16. **RATING (score level)** - What is the level of certainty or confidence that the performance measure scores are reliable?

    TIPS: Consider the following: Is the test sample adequate to generalize for widespread implementation?

    Do the results demonstrate sufficient reliability so that differences in performance can be identified?

    ☐High (go to Question #8)

    ☒Moderate (go to Question #8)

    ☐Low (please explain below then go to Question #7)

    Reliability for all hospital was low (95% interval of (0.42,0.58) and moderate for hospitals with at least 10 procedures performed (0.76, 0.910. Recommend that the developer consider limiting the entities to those with at least 10 procedures.

17. Was other reliability testing reported?

    ☒Yes (go to Question #8)

    ☐No (rate Question #11: OVERALL RELIABILITY as LOW and proceed to the VALIDITY SECTION)

18. Was reliability testing conducted with patient-level data elements that are used to construct the performance measure?

    TIPS: Prior reliability studies of the same data elements may be submitted; if comparing abstraction to "authoritative source/gold standard" see Validity Section Question #15)

    ☐Yes (go to Question #9)

    ☒No (if there is score-level testing, rate Question #11: OVERALL RELIABILITY based on score-level rating from Question #6; otherwise, rate Question #11: OVERALL RELIABILITY as INSUFFICIENT. Then proceed to the VALIDITY SECTION)

    Developer mentions testing elements via random review (mentioned in Section 2b1.2 under validity). I think the audit is actually testing reliability and should be reported as a Kappa statistic. If the measure developer is treating the auditing firm as the 'gold standard', then this could be considered a validity measure. As reported, it is hard to determine how they are treating the audit – other than including the description in the 'validity' section of the report.

19. Was the method described and appropriate for assessing the reliability of ALL critical data elements?

    TIPS: For example: inter-abstractor agreement (ICC, Kappa); other accepted method with description of how it assesses reliability of the data elements

    Answer no if: only assessed percent agreement; did not assess separately for all data elements (at least numerator, denominator, exclusions)

☐Yes (go to Question #10)

☐No (if no, please explain below and rate Question #10 as INSUFFICIENT)

20. **RATING (data element)** – Based on the reliability statistic and scope of testing (number and representativeness of patients and entities), what is the level of certainty or confidence that the data used in the measure are reliable?

TIPS: Consider the following: Is the test sample adequate to generalize for widespread implementation? Can data elements be collected consistently?

☐Moderate (if score-level testing was NOT conducted, rate Question #11: OVERALL RELIABILITY as MODERATE)

☐Low (if score-level testing was NOT conducted, rate Question #11: OVERALL RELIABILITY as LOW)

☐Insufficient (go to Question #11)

11. OVERALL RELIABILITY RATING

OVERALL RATING OF RELIABILITY taking into account precision of specifications and all testing results:

☐High (NOTE: Can be HIGH only if score-level testing has been conducted)

☒Moderate (NOTE: Moderate is the highest eligible rating if score-level testing has not been conducted)

☐Low (please explain below) [NOTE: Should rate LOW if you believe specifications are NOT precise, unambiguous, and complete]

☐Insufficient (please explain below) [NOTE: For most measure types, testing at both the score level and the data element level is not required]

Reliability measure is heavily dependent on the number of procedures per hospital – not unexpected, but developer should consider implementing a lower bound on the number of observations per entity for application of the measure.

VALIDITY

ASSESSMENT OF THREATS TO VALIDITY

17. Were all potential threats to validity that are relevant to the measure empirically assessed?

TIPS: Threats to validity include: exclusions; need for risk adjustment; Able to identify statistically significant and meaningful differences; multiple sets of specifications; missing data/nonresponse.

☒Yes (go to Question #2)

☐No (please explain below and go to Question #2) [NOTE that even if ***non-assessment of applicable threats should result in an overall INSUFFICENT rating for validity***, we still want you to look at the testing results]

18. Analysis of potential threats to validity: Any concerns with measure exclusions?

TIPS: Consider the following: Are the exclusions consistent with the evidence? Are any patients or patient groups inappropriately excluded from the measure? Are the exclusions/exceptions of sufficient frequency and variation across providers to be needed (and outweigh the data collection burden)? If patient preference (e.g., informed decisionmaking) is a basis for exclusion, does it impact performance and if yes, is the measure specified so that the information about patient preference and the effect on the measure is transparent?

☐Yes (please explain below then go to Question #3)

☒No (go to Question #3)

☐Not applicable (i.e., there are no exclusions specified for the measure; go to Question #3)

Small number of exclusions, but no assessment of the impact. 2b2.3 does not mention the number of patients with missing 'discharge mortality status' – since this is one of the measured outcomes, missing values may cause measurement bias.

19. Analysis of potential threats to validity: Risk-adjustment (applies to all outcome, cost, and resource use measures; may also apply to other types of measure)

☐Not applicable (e.g., structure or process measure that is not risk-adjusted; go to Question #4)

a. Is a conceptual rationale for social risk factors included? ☐Yes ☒No

b. Are social risk factors included in risk model? ☐Yes ☒No

STS mentions that dual eligibility might serve as a proxy for social risk. I agree with this premise, but they did not include it as a risk adjustor. No explanation other than stating "However, this information is not presently included in STS data analysis nor as a basis for stratification in STS measures."

c. Any concerns regarding the risk-adjustment approach?

*TIPS: Consider the following: If a justification for **not risk adjusting** is provided, is there any evidence that contradicts the developer's rationale and analysis? If the developer asserts there is **no conceptual basis** for adjusting this measure for social risk factors, do you agree with the rationale? **If risk adjusted**: Are the candidate and final variables included in the risk adjustment model adequately described for the measure to be implemented? Are the candidate and final variables included in the risk adjustment model adequately described for the measure to be implemented? Are all of the risk adjustment variables present at the start of care (if not, do you agree with the rationale)? If social risk factors are not included in the risk-adjustment approach, do you agree with the developer's decision? Is an appropriate risk-adjustment strategy included in the measure (e.g., adequate model discrimination and calibration)? Are all statistical model specifications included, including a "clinical model only" if social risk factors are included in the final model?*

☒Yes (please explain below then go to Question #4)

☐No (go to Question #4)

There is no mention of how potential multi-collinearity among the risk adjustors was either assessed or addressed. This combined with the inclusion of risk adjustment variables with no statistical evidence of predictive value compromises the value of the risk-adjustment approach. For example, none of the pathological stage variables have an odds ratio with a confidence interval excluding 1.0.

Section 2b3.8 refers to a risk decile plot, but and ROC curve is displayed instead.

20. Analysis of potential threats to validity: Any concerns regarding ability to identify meaningful differences in performance or overall poor performance?

☒Yes (please explain below then go to Question #5)

☐No (go to Question #5)

Small sample hospitals may compromise the stability of the risk model. All selected variables were included in the model – many have odds ratios with CI that cover 1.0 and are not statistically significant.

21. Analysis of potential threats to validity: Any concerns regarding comparability of results if multiple data sources or methods are specified?

☐Yes (please explain below then go to Question #6)

☐No (go to Question #6)

☒Not applicable (go to Question #6)

22. Analysis of potential threats to validity: Any concerns regarding missing data?

☐Yes (please explain below then go to Question #7)

☒No (go to Question #7)

ASSESSMENT OF MEASURE TESTING

23. Was empirical validity testing conducted using the measure as specified and appropriate statistical test?

    Answer no if: face validity; only refer to clinical evidence; only descriptive statistics; only describe process for data management/cleaning/computer programming; testing does not match measure specifications (i.e. data, eMeasure, level, setting, patients).

    ☒Yes (go to Question #10) [NOTE: If appropriate empirical testing has been conducted, then evaluation of face validity is not necessary. Go to Question #8 **only if** there is insufficient information provided to evaluate data element and score-level testing.]

    ☐No (please explain below then go to Question #8)

    Agreement rates with auditor reported – see Question #8 under reliability.

24. Was face validity systematically assessed by recognized experts to determine agreement on whether the computed performance measure score from the measure as specified can be used to distinguish good and poor quality?

    TIPS: Answer no if: focused on data element accuracy/availability/feasibility/other topics; the degree of consensus and any areas of disagreement not provided/discussed.

    ☐Yes (go to Question #9)

    ☐No (please explain below and rate Question #17: OVERALL VALIDITY as INSUFFICIENT)

25. **RATING (face validity)** - Do the face validity testing results indicate substantial agreement that the performance measure score from the measure as specified can be used to distinguish quality AND potential threats to validity are not a problem, OR are adequately addressed so results are not biased?

    ☐Yes (if a NEW measure, rate Question #17: OVERALL VALIDITY as MODERATE)

    ☐ Yes (if a MAINTENANCE measure, do you agree with the justification for not

    conducting empirical testing? If no, rate Question #17: OVERALL VALIDITY as

INSUFFICIENT; otherwise, rate Question #17: OVERALL VALIDITY as MODERATE)

    ☐No (please explain below and rate Question #17: OVERALL VALIDITY AS LOW)

26. Was validity testing conducted with computed performance measure scores for each measured entity?

    TIPS: Answer no if: one overall score for all patients in sample used for testing patient-level data.

    ☐Yes (go to Question #11)

    ☒No (please explain below and go to Question #13)

    I am interpreting 'measured entity' to be hospital for this measure. I do not see a hospital level assessment of validity on the computed score. I do see those results for the individual data elements for a sample of the measured entities.

27. Was the method described and appropriate for assessing conceptually and theoretically sound hypothesized relationships?

    TIPS: For example: correlation of the performance measure score on this measure and other performance measures; differences in performance scores between groups known to differ on quality; other accepted method with description of how it assesses validity of the performance score

    ☐Yes (go to Question #12)

    ☐No (please explain below, rate Question #12 as INSUFFICIENT and then go to Question #14)

28. **RATING (measure score)** - Based on the measure score results (significance, strength) and scope of testing (number of measured entities and representativeness) and analysis of potential threats, what is the level of certainty or confidence that the performance measure scores are a valid indicator of quality?

☐High (go to Question #14)

☐Moderate (go to Question #14)

☐Low (please explain below then go to Question #13)

☐Insufficient

29. Was other validity testing reported?

☒Yes (go to Question #14)

☐No (please explain below and rate Question #17: OVERALL VALIDITY as LOW)

30. Was validity testing conducted with patient-level data elements?

TIPS: Prior validity studies of the same data elements may be submitted

☒Yes (go to Question #15)

☐No (please explain below and rate Question #17: OVERALL VALIDITY as INSUFFICIENT if no score-level testing was conducted, otherwise, rate Question #17: OVERALL VALIDITY based on score-level rating from Question #12)

Reported agreement rates with quality audits as validity measure. Agreement rates are high for most data elements (95% +)

31. Was the method described and appropriate for assessing the accuracy of ALL critical data elements? *NOTE that data element validation from the literature is acceptable.*

TIPS: For example: Data validity/accuracy as compared to authoritative source- sensitivity, specificity, PPV, NPV; other accepted method with description of how it assesses validity of the data elements.

Answer No if: only assessed percent agreement; did not assess separately for all data elements (at least numerator, denominator, exclusions)

☒Yes (go to Question #16)

☐No (please explain below and rate Question #16 as INSUFFICIENT)

32. **RATING (data element)** - Based on the data element testing results (significance, strength) and scope of testing (number and representativeness of patients and entities) and analysis of potential threats, what is the level of certainty or confidence that the data used in the measure are valid?

☒Moderate (if score-level testing was NOT conducted, rate Question #17: OVERALL VALIDITY as MODERATE)

☐Low (please explain below) (if score-level testing was NOT conducted, rate Question #17: OVERALL VALIDITY as LOW)

☐Insufficient (go to Question #17)

17. OVERALL VALIDITY RATING

OVERALL RATING OF VALIDITY taking into account the results and scope of all testing and analysis of potential threats.

☒High (NOTE: Can be HIGH only if score-level testing has been conducted)

☐Moderate (NOTE: Moderate is the highest eligible rating if score-level testing has NOT been conducted)

☐Low (please explain below) [NOTE: Should rate LOW if you believe that there are threats to validity and/or threats to validity were not assessed]

☐Insufficient (if insufficient, please explain below) [NOTE: For most measure types, testing at both the score level and the data element level is not required] [NOTE: If rating is INSUFFICIENT for all empirical testing, then go back to Question #8 and evaluate any face validity that was conducted, then reconsider this overall rating.]

FOR COMPOSITE MEASURES ONLY: Empirical analyses to support composite construction

What is the level of certainty or confidence that the empirical analysis demonstrates that the component measures add value to the composite and that the aggregation and weighting rules are consistent with the quality construct?

TIPS: Consider the following: Do the component measures fit the quality construct? Are the objectives of parsimony and simplicity achieved while supporting the quality construct?

☐High

☐Moderate

☐Low (please explain below)

☐Insufficient (please explain below)

Evaluation C: Scientific Acceptability

Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. **Measures must be judged to meet the subcriteria for both reliability and validity to pass this criterion.**

Instructions:

- Please complete this form for each measure you are evaluating.
- Please pay close attention to the skip logic directions.
- If you are unable to check a box, please highlight or shade the box for your response.
- You must answer the "overall rating" item for both Reliability and Validity. Also, be sure to answer the composite measure question at the end of the form if your measure is a composite.
- We have provided TIPS to help you answer the questions.
- We've designed this form to try to minimize the amount of writing that you have to do. That said, ***it is critical that you explain your thinking/rationale if you check boxes where we ask for an explanation*** (because this is a Word document, you can just add your explanation below the checkbox)**.** Feel free to add additional explanation, even if an explanation is not requested (but please type this underneath the appropriate checkbox).
- This form is based on Algorithms 2 and 3 in the Measure Evaluation Criteria and Guidance document (see pages 18-24). These algorithms provide guidance to help you rate the Reliability and Validity subcriteria. ***We ask that you refer to this document when you are evaluating your measures***.
- Please contact Methods Panel staff if you have questions (methodspanel@qualityforum.org).

Measure Number: 1790

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

RELIABILITY

21. Are submitted specifications precise, unambiguous, and complete so that they can be consistently implemented?

    NOTE: NQF staff will conduct a separate, more technical, check of eMeasure (eCQM) specifications, value sets, logic, and feasibility, so no need to consider these in your evaluation.

    TIPS: Consider the following: Are all the data elements clearly defined? Are all appropriate codes included? Is the logic or calculation algorithm clear? Is it likely this measure can be consistently implemented?

    ☒Yes (go to Question #2)

    ☐No (please explain below, and go to Question #2) NOTE that even though ***non-precise specifications should result in an overall LOW rating for reliability***, we still want you to look at the testing results.

22. Was empirical reliability testing (at the data element or measure score level) conducted using statistical tests with the measure as specified?

TIPS: Check the 2nd "NO" box below if: only descriptive statistics provided; only describes process for data management/cleaning/computer programming; testing does not match measure specifications (i.e. data, eMeasure, level of analysis, patients)

☒Yes (go to Question #4)

☐No, there is reliability testing information, but *not* using statistical tests and/or not for the measure as specified OR there is no reliability testing (please explain below then go to Question #3)

23. Was **empirical VALIDITY testing** of patient-level data conducted?

    ☐Yes (use your rating from data element validity testing – Question #16- under Validity Section)

    ☐No (please explain below and rate Question #11: OVERALL RELIABILITY as INSUFFICIENT and proceed to the VALIDITY SECTION)

24. Was reliability testing conducted with computed performance measure scores for each measured entity?

    TIPS: Answer no if: only one overall score for all patients in sample used for testing patient-level data

    ☒Yes (go to Question #5)

    ☐No (go to Question #8)

25. Was the method described and appropriate for assessing the proportion of variability due to real differences among measured entities? *NOTE: If multiple methods used, at least one must be appropriate.*

    TIPS: Examples of appropriate methods include signal-to-noise analysis (e.g. Adams/RAND tutorial); random split-half correlation; other accepted method with description of how it assesses reliability of the performance score.

    ☒Yes (go to Question #6) (The method description could be made better by providing reference(s), slightly more information about Bayesian estimation of the true value. I think the denominator of the equation is missing a superscript. In addition, on page 5, 3rd row from the bottom, one notation is off, theta should be rho.)

    ☐No (please explain below then go to Question #8)

26. **RATING (score level)** - What is the level of certainty or confidence that the performance measure scores are reliable?

    TIPS: Consider the following: Is the test sample adequate to generalize for widespread implementation?

    Do the results demonstrate sufficient reliability so that differences in performance can be identified?

    ☐High (go to Question #8)

    ☒Moderate (go to Question #8)

    ☐Low (please explain below then go to Question #7)

27. Was other reliability testing reported?

    ☐Yes (go to Question #8)

    ☐No (rate Question #11: OVERALL RELIABILITY as LOW and proceed to the VALIDITY SECTION)

28. Was reliability testing conducted with patient-level data elements that are used to construct the performance measure?

    TIPS: Prior reliability studies of the same data elements may be submitted; if comparing abstraction to "authoritative source/gold standard" see Validity Section Question #15)

    ☐Yes (go to Question #9)

    ☒No (if there is score-level testing, rate Question #11: OVERALL RELIABILITY based on score-level rating from Question #6; otherwise, rate Question #11: OVERALL RELIABILITY as INSUFFICIENT. Then proceed to the VALIDITY SECTION)

29. Was the method described and appropriate for assessing the reliability of ALL critical data elements?

TIPS: For example: inter-abstractor agreement (ICC, Kappa); other accepted method with description of how it assesses reliability of the data elements

Answer no if: only assessed percent agreement; did not assess separately for all data elements (at least numerator, denominator, exclusions)

☐Yes (go to Question #10)

☐No (if no, please explain below and rate Question #10 as INSUFFICIENT)

30. **RATING (data element)** – Based on the reliability statistic and scope of testing (number and representativeness of patients and entities), what is the level of certainty or confidence that the data used in the measure are reliable?

TIPS: Consider the following: Is the test sample adequate to generalize for widespread implementation? Can data elements be collected consistently?

☐Moderate (if score-level testing was NOT conducted, rate Question #11: OVERALL RELIABILITY as MODERATE)

☐Low (if score-level testing was NOT conducted, rate Question #11: OVERALL RELIABILITY as LOW)

☐Insufficient (go to Question #11)

11. OVERALL RELIABILITY RATING

OVERALL RATING OF RELIABILITY taking into account precision of specifications and all testing results:

☐High (NOTE: Can be HIGH only if score-level testing has been conducted)

☒Moderate (NOTE: Moderate is the highest eligible rating if score-level testing has not been conducted)

☐Low (please explain below) [NOTE: Should rate LOW if you believe specifications are NOT precise, unambiguous, and complete]

☐Insufficient (please explain below) [NOTE: For most measure types, testing at both the score level and the data element level is not required]

VALIDITY

ASSESSMENT OF THREATS TO VALIDITY

33. Were all potential threats to validity that are relevant to the measure empirically assessed?

TIPS: Threats to validity include: exclusions; need for risk adjustment; Able to identify statistically significant and meaningful differences; multiple sets of specifications; missing data/nonresponse.

☒Yes (go to Question #2)

☐No (please explain below and go to Question #2) [NOTE that even if ***non-assessment of applicable threats should result in an overall INSUFFICENT rating for validity***, we still want you to look at the testing results]

34. Analysis of potential threats to validity: Any concerns with measure exclusions?

TIPS: Consider the following: Are the exclusions consistent with the evidence? Are any patients or patient groups inappropriately excluded from the measure? Are the exclusions/exceptions of sufficient frequency and variation across providers to be needed (and outweigh the data collection burden)? If patient preference (e.g., informed decisionmaking) is a basis for exclusion, does it impact performance and if yes, is the measure specified so that the information about patient preference and the effect on the measure is transparent?

☐Yes (please explain below then go to Question #3)

☒No (go to Question #3) (It would be helpful if the developer quantifies the exclusion rates for various exclusion criteria. In the attached paper, it did describe the overall exclusion rate, but it would be better if they provide criterion specific information in the testing form.)

☐Not applicable (i.e., there are no exclusions specified for the measure; go to Question #3)

35. Analysis of potential threats to validity: Risk-adjustment (applies to all outcome, cost, and resource use measures; may also apply to other types of measure)

☐Not applicable (e.g., structure or process measure that is not risk-adjusted; go to Question #4)

a. Is a conceptual rationale for social risk factors included? ☐Yes ☒No (The developer did not conduct any analysis related to social risk factors and did not provide any conceptual rationale for social risk factors. They could consider using other information as a proxy for patient social risk factors, for example, using insurance information to identify dual eligible patients as they mentioned in their testing form.)

b. Are social risk factors included in risk model? ☐Yes ☒No

c. Any concerns regarding the risk-adjustment approach?

*TIPS: Consider the following: If a justification for **not risk adjusting** is provided, is there any evidence that contradicts the developer's rationale and analysis? If the developer asserts there is **no conceptual basis** for adjusting this measure for social risk factors, do you agree with the rationale? **If risk adjusted**: Are the candidate and final variables included in the risk adjustment model adequately described for the measure to be implemented? Are the candidate and final variables included in the risk adjustment model adequately described for the measure to be implemented? Are all of the risk adjustment variables present at the start of care (if not, do you agree with the rationale)? If social risk factors are not included in the risk-adjustment approach, do you agree with the developer's decision? Is an appropriate risk-adjustment strategy included in the measure (e.g., adequate model discrimination and calibration)? Are all statistical model specifications included, including a "clinical model only" if social risk factors are included in the final model?*

☒Yes (please explain below then go to Question #4) (In general, it is fine. But it would be very helpful to have an external validation of the risk adjustment model. That is, the model developed based on the development sample works similarly well in a validation sample.)

☐No (go to Question #4)

36. Analysis of potential threats to validity: Any concerns regarding ability to identify meaningful differences in performance or overall poor performance?

☐Yes (please explain below then go to Question #5)

☒No (go to Question #5)

37. Analysis of potential threats to validity: Any concerns regarding comparability of results if multiple data sources or methods are specified?

☐Yes (please explain below then go to Question #6)

☒No (go to Question #6)

☐Not applicable (go to Question #6)

38. Analysis of potential threats to validity: Any concerns regarding missing data?

☐Yes (please explain below then go to Question #7)

☒No (go to Question #7) (The potential concern is if participating facilities submit all their cases, that is, not selectively submit their cases. The developer mentioned (page 8) "there was consistent agreement across all participants for data completeness." Does this really mean 100% for all sites? If yes, it is good to know; if not, then that needs to be quantified.)

ASSESSMENT OF MEASURE TESTING

39. Was empirical validity testing conducted using the measure as specified and appropriate statistical test?

    Answer no if: face validity; only refer to clinical evidence; only descriptive statistics; only describe process for data management/cleaning/computer programming; testing does not match measure specifications (i.e. data, eMeasure, level, setting, patients).

    ☒Yes (go to Question #10) [NOTE: If appropriate empirical testing has been conducted, then evaluation of face validity is not necessary. Go to Question #8 **only if** there is insufficient information provided to evaluate data element and score-level testing.] (Measure testing was done at the hospital level only although the developer checked both hospital and group practice (section 1.4), the developer should uncheck "group practice".).

    ☐No (please explain below then go to Question #8)

40. Was face validity systematically assessed by recognized experts to determine agreement on whether the computed performance measure score from the measure as specified can be used to distinguish good and poor quality?

    TIPS: Answer no if: focused on data element accuracy/availability/feasibility/other topics; the degree of consensus and any areas of disagreement not provided/discussed.

    ☐Yes (go to Question #9)

    ☐No (please explain below and rate Question #17: OVERALL VALIDITY as INSUFFICIENT)

41. **RATING (face validity)** - Do the face validity testing results indicate substantial agreement that the performance measure score from the measure as specified can be used to distinguish quality AND potential threats to validity are not a problem, OR are adequately addressed so results are not biased?

    ☐Yes (if a NEW measure, rate Question #17: OVERALL VALIDITY as MODERATE)

    ☐ Yes (if a MAINTENANCE measure, do you agree with the justification for not conducting empirical testing? If no, rate Question #17: OVERALL VALIDITY as INSUFFICIENT; otherwise, rate Question #17: OVERALL VALIDITY as MODERATE)

    ☐No (please explain below and rate Question #17: OVERALL VALIDITY AS LOW)

42. Was validity testing conducted with computed performance measure scores for each measured entity?

    TIPS: Answer no if: one overall score for all patients in sample used for testing patient-level data.

    ☐Yes (go to Question #11)

    ☒No (please explain below and go to Question #13) (The developer checked "Empirical validity testing (page 7) leaving both critical data elements and performance measure score unchecked.)

43. Was the method described and appropriate for assessing conceptually and theoretically sound hypothesized relationships?

    TIPS: For example: correlation of the performance measure score on this measure and other performance measures; differences in performance scores between groups known to differ on quality; other accepted method with description of how it assesses validity of the performance score

    ☐Yes (go to Question #12)

    ☐No (please explain below, rate Question #12 as INSUFFICIENT and then go to Question #14)

44. **RATING (measure score)** - Based on the measure score results (significance, strength) and scope of testing (number of measured entities and representativeness) and analysis of potential threats, what is the level of certainty or confidence that the performance measure scores are a valid indicator of quality?

    ☐High (go to Question #14)

☐Moderate (go to Question #14)

☐Low (please explain below then go to Question #13)

☐Insufficient

45. Was other validity testing reported?

☒Yes (go to Question #14)

☐No (please explain below and rate Question #17: OVERALL VALIDITY as LOW)

46. Was validity testing conducted with patient-level data elements?

TIPS: Prior validity studies of the same data elements may be submitted

☒Yes (go to Question #15)

☐No (please explain below and rate Question #17: OVERALL VALIDITY as INSUFFICIENT if no score-level testing was conducted, otherwise, rate Question #17: OVERALL VALIDITY based on score-level rating from Question #12)

47. Was the method described and appropriate for assessing the accuracy of ALL critical data elements? *NOTE that data element validation from the literature is acceptable.*

TIPS: For example: Data validity/accuracy as compared to authoritative source- sensitivity, specificity, PPV, NPV; other accepted method with description of how it assesses validity of the data elements.

Answer No if: only assessed percent agreement; did not assess separately for all data elements (at least numerator, denominator, exclusions)

☒Yes (go to Question #16)

(A. Concerns for two critical data elements, one is "FEV1 Predicted" (page 8), this is a risk adjustment variable. The agreement rate for this variable is only 76.33%. Another is "Status 30 Days after surgery" (page 9), even though the agreement rate is 92.40%, given that this is a very important endpoint that should have little ambiguity, basically 30-day mortality, I would hope the agreement rate for this variable is higher. If the agreement varies across hospitals, then it would be even more concerning.

(B. Although the developer did not provide kappa statistic as required due to lack of patient level data, in this case, having percent agreement is sufficiently informative. In fact, in some situations, reporting kappa statistic without percent agreement is worse than reporting percent agreement without kappa. It is known that in some situations the observed proportion of agreement "can be paradoxically altered by the chance-corrected ratio that creates kappa as an index of concordance." (See "High agreement but low kappa: I. The problems of two paradoxes", Feinstein & Cicchetti))

☐No (please explain below and rate Question #16 as INSUFFICIENT)

48. **RATING (data element)** - Based on the data element testing results (significance, strength) and scope of testing (number and representativeness of patients and entities) and analysis of potential threats, what is the level of certainty or confidence that the data used in the measure are valid?

☒Moderate (if score-level testing was NOT conducted, rate Question #17: OVERALL VALIDITY as MODERATE) (I don't see score level validity testing, some concerns about two critical data elements.)

☐Low (please explain below) (if score-level testing was NOT conducted, rate Question #17: OVERALL VALIDITY as LOW)

☐Insufficient (go to Question #17)

17. OVERALL VALIDITY RATING

OVERALL RATING OF VALIDITY taking into account the results and scope of all testing and analysis of potential threats.

☐High (NOTE: Can be HIGH only if score-level testing has been conducted)

☒Moderate (NOTE: Moderate is the highest eligible rating if score-level testing has NOT been conducted)

☐Low (please explain below) [NOTE: Should rate LOW if you believe that there are threats to validity and/or threats to validity were not assessed]

☐Insufficient (if insufficient, please explain below) [NOTE: For most measure types, testing at both the score level and the data element level is not required] [NOTE: If rating is INSUFFICIENT for all empirical testing, then go back to Question #8 and evaluate any face validity that was conducted, then reconsider this overall rating.]

FOR COMPOSITE MEASURES ONLY: Empirical analyses to support composite construction

What is the level of certainty or confidence that the empirical analysis demonstrates that the component measures add value to the composite and that the aggregation and weighting rules are consistent with the quality construct?

TIPS: Consider the following: Do the component measures fit the quality construct? Are the objectives of parsimony and simplicity achieved while supporting the quality construct?

☐High

☐Moderate

☐Low (please explain below)

☐Insufficient (please explain below)

Committee Pre-evaluation Comments: Criteria 2: Scientific Acceptability of Measure Properties (including all 2a, 2b, and 2c)

2a1. Reliability Specifications

**None

**High Reliability

**Specifications are clear – it would be interesting to see reporting for rural versus urban and low volume versus high volume centers

** The STS database has demonstrated well-defined data elements that continue to be refined. Site audits are performed to confirm data reliability. Sophisticated risk adjustment algorithms have been validated for their reliability.

2a2. Reliability Testing

**No

**No

**No

2b1. Validity Testing

**No

**Valid

**No

**The validity of the STS database has been evaluated in detail.

2b2-3. Other threats to validity

**Appropriately risk adjusted, and not unduly burdensome for those already participating in the STS registry. It will be nearly impossible to use this measure otherwise, however.

**Risk adjusted w adequate addressing of GTSDB lack of social risk factors

** I have no issues

**Appropriate exclusions have been made, primarily for low volume but high risk procedures that otherwise may skew the results.

Criterion 3. Feasibility

Maintenance measures – no change in emphasis – implementation issues may be more prominent

3. Feasibility is the extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

- The developer reports that data elements are generated by and used by healthcare personnel during the provision of care. Data are also coded by someone other than the person obtaining the original information and abstracted from a record by someone other than the person obtaining the original information.
- The developer provided the costs associated with the STS registry.

Questions for the Committee:

- o *Are the required data elements routinely generated and used during care delivery?*
- o *Are the required data elements available in electronic form, e.g., EHR or other electronic sources?*

Preliminary rating for feasibility: ☒ **High** ☐ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee Pre-evaluation Comments: Criteria 3: Feasibility

3a. Feasibility

**STS registry participation is expensive and burdensome, but offers a substantial return on investment.
**My only concern is penetrance of STS GTSDB. Historically the GTSDB is comprised of the highest TS performers.
**Participation in STS is costly, however there is widespread participation among facilities performing these procedures
**Although cost and other resources are required to participate in the STS registry, all but a few centers in the United States are currently participating.

Criterion 4: Usability and Use

Maintenance measures – increased emphasis – much greater focus on measure use and usefulness, including both impact/improvement and unintended consequences

4a. Use (4a1. Accountability and Transparency; 4a2. Feedback on measure)

4a. Use evaluate the extent to which audiences (e.g., consumers, purchasers, providers, policymakers) use or could use performance results for both accountability and performance improvement activities.

4a.1. Accountability and Transparency. Performance results are used in at least one accountability application within three years after initial endorsement and are publicly reported within six years after initial endorsement (or the data on performance results are available). If not in use at the time of initial endorsement, then a credible plan for implementation within the specified timeframes is provided.

Current uses of the measure

Publicly reported? ☒ **Yes** ☐ **No**

Current use in an accountability program? ☒ **Yes** ☐ **No** ☐ **UNCLEAR**

OR

Planned use in an accountability program? ☐ **Yes** ☐ **No**

Accountability program details

- The measure results are shared with participants in the STS General Thoracic Surgery Database (GTSD) for quality improvement purposes. In addition, the developer reports active promotion of STS measures through the STS Public Reporting Task force. The task force develops public report cards that are consumer centric.

4a.2. Feedback on the measure by those being measured or others. Three criteria demonstrate feedback: 1) those being measured have been given performance results or data, as well as assistance with interpreting the measure results and data; 2) those being measured and other users have been given an opportunity to provide feedback on the

measure performance or implementation; 3) this feedback has been considered when changes are incorporated into the measure

Feedback on the measure by those being measured or others:

- The developer states that STS surgeon members have expressed interest in real-time, online data updates which led to the development of a general thoracic dashboard. The dashboard is scheduled for launch in 2018.

Additional Feedback:

- The developer reports that surgeons on the STS General Thoracic Surgery Task Force meet periodically to discuss participant reports and discuss enhancements to the GTS database. Additions and clarifications to the data collection form and the content/format of participant reports are discussed and implemented as appropriate.
- The developer noted that the report *Data Analyses of the Society of Thoracic Surgeons General Thoracic Surgery Database* displays results for Combined Morbidity/Mortality for Pulmonary Resections. These data are shown at the participant level and in comparison to the 25th, 50th, and 75th percentiles across all participants in the STS database. The data area also shared to participants semi-annually.

***Questions for the Committee**:*

- o *How have (or can) the performance results be used to further the goal of high-quality, efficient healthcare?*
- o *How has the measure been vetted in real-world settings by those being measured or others?*

Preliminary rating for Use: ☒ **Pass** ☐ **No Pass**

4b. Usability (4a1. Improvement; 4a2. Benefits of measure)

4b. Usability evaluate the extent to which audiences (e.g., consumers, purchasers, providers, policymakers) use or could use performance results for both accountability and performance improvement activities.

4b.1 Improvement. Progress toward achieving the goal of high-quality, efficient healthcare for individuals or populations is demonstrated.

Improvement results

- The developer reports that operative mortality in the STS General Thoracic Surgery Database (GTSD) decreased from 2.2% (from 2002-2008) to 1.4% (from 2012-2014). Further, when data from the GTSD were compared with the Nationwide Inpatient Sample database from 2002 to 2008, patients in the GTSD had lower unadjusted mortality rates, median length of stay, and lower pulmonary complication rates for lobectomy.

4b2. Benefits vs. harms. Benefits of the performance measure in facilitating progress toward achieving high-quality, efficient healthcare for individuals or populations outweigh evidence of unintended negative consequences to individuals or populations (if such evidence exists).

Unexpected findings (positive or negative) during implementation

- The developer reports they are unaware of any unexpected findings associated with the implementation of this measure.

Potential harms

- The developer reports that the rate of major morbidity has increased from 8.6% to 9.1% from 2002 to 2008 which is potentially explained by more complete coding of complications by data abstractors and inclusion of unexpected return to the operating room for any reason.

***Questions for the Committee**:*

- o *How can the performance results be used to further the goal of high-quality, efficient healthcare?*
- o *Do the benefits of the measure outweigh any potential unintended consequences?*

Preliminary rating for Usability : ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee Pre-evaluation Comments: Criteria 4: Usability and Use

4a1. Use
**Closely followed by facilities and surgeons.
**Transparent
**No issues
**Participating institutions receive risk-adjusted reports has to their performance. The STS also has a public reporting task force that develops report cards for the consumer.
4b1. Usability
**Public reporting will increase attention to performance
**Usable
**No issues other than the cost to obtain the clinical data
**The institution level reports are designed to guide process improvement initiatives.

Criterion 5: Related and Competing Measures

Related or competing measures

- 3294 STS Lobectomy for Lung Cancer Composite Score
- The developer notes that NQF 1790 is related conceptually to 3294 and that the numerators for both measures include the same list of postoperative complications, but the outcomes for the Lobectomy Composite measure are grouped into two domains (operative mortality and major complications) and the measure is structured to provide general thoracic surgeons with a "star rating."
- Measure #1790 includes a broader range of lung resection procedures than the Lobectomy Composite, and therefore includes a larger number of cases and potentially provides performance data to more general thoracic surgeons.

Harmonization

- The developer reports that NQF 1790 and 3294 are harmonized to the extent possible.

Committee Pre-evaluation Comments: Criterion 5: Related and Competing Measures

Public and Member Comments

Comments and Member Support/Non-Support Submitted as of: January 18, 2018

- No NQF members have submitted support/non-support choices as of this date. No comments have been submitted as of this date.

1. Evidence and Performance Gap – Importance to Measure and Report

Extent to which the specific measure focus is evidence-based, important to making significant gains in healthcare quality, and improving health outcomes for a specific high-priority (high-impact) aspect of healthcare where there is variation in or overall less-than-optimal performance. ***Measures must be judged to meet all sub criteria to pass this criterion and be evaluated against the remaining criteria.***

1a. Evidence to Support the Measure Focus – See attached Evidence Submission Form

NQF_evidence_attachment_STS-1790-111517-v2.docx

1a.1 For Maintenance of Endorsement: Is there new evidence about the measure since the last update/submission? Do not remove any existing information. If there have been any changes to evidence, the Committee will consider the new evidence. Please use the most current version of the evidence attachment (v7.1). Please use red font to indicate updated evidence.

No

1a Evidence (subcriterion 1a)

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

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

IF the measure is a component in a composite performance measure, provide the title of the Composite Measure here:

Date of Submission: 11/15/2017

Instructions

- *Complete 1a.1 and 1a.2 for all measures. If instrument-based measure, complete 1a.3.*
- *Complete **EITHER 1a.2, 1a.3 or 1a.4** as applicable for the type of measure and evidence.*
- *For composite performance measures:*
  - *A separate evidence form is required for each component measure unless several components were studied together.*
  - *If a component measure is submitted as an individual performance measure, attach the evidence form to the individual measure submission.*
- All information needed to demonstrate meeting the evidence subcriterion (1a) must be in this form. An appendix of *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.
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.

Note: The information provided in this form is intended to aid the Standing Committee and other stakeholders in understanding to what degree the evidence for this measure meets NQF's evaluation criteria.

1a. Evidence to Support the Measure Focus

The measure focus is evidence-based, demonstrated as follows:

- Outcome: [3] Empirical data demonstrate a relationship between the outcome and at least one healthcare structure, process, intervention, or service. If not available, wide variation in performance can be used as evidence, assuming the data are from a robust number of providers and results are not subject to systematic bias.
- Intermediate clinical outcome: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured intermediate clinical outcome leads to a desired health outcome.

- Process: [5] a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured process leads to a desired health outcome.
- Structure: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured structure leads to a desired health outcome.
- Efficiency: [6] evidence not required for the resource use component.
- For measures derived from patient reports, evidence should demonstrate that the target population values the measured outcome, process, or structure and finds it meaningful.
- Process measures incorporating Appropriate Use Criteria: See NQF's guidance for evidence for measures, in general; guidance for measures specifically based on clinical practice guidelines apply as well.

Notes

3. Generally, rare event outcomes do not provide adequate information for improvement or discrimination; however, serious reportable events that are compared to zero are appropriate outcomes for public reporting and quality improvement.

4. The preferred systems for grading the evidence are the Grading of Recommendations, Assessment, Development and Evaluation (GRADE) guidelines and/or modified GRADE.

5. Clinical care processes typically include multiple steps: assess → identify problem/potential problem → choose/plan intervention (with patient input) → provide intervention → evaluate impact on health status. If the measure focus is one step in such a multistep process, the step with the strongest evidence for the link to the desired outcome should be selected as the focus of measurement. Note: A measure focused only on collecting PROM data is not a PRO-PM.

6. Measures of efficiency combine the concepts of resource use and quality (see NQF's Measurement Framework: Evaluating Efficiency Across Episodes of Care; AQA Principles of Efficiency Measures).

1a.1.This is a measure of: (*should be consistent with type of measure entered in De.1*)

Outcome

☒ Outcome: Postoperative complications: reintubation, need for tracheostomy, initial ventilator support > 48 hours, ARDS, pneumonia, pulmonary embolus, bronchopleural fistula, unexpected return to the operating room, myocardial infarction or operative mortality (death during the index hospitalization, regardless of timing, or within 30 days, regardless of location).

☐Patient-reported outcome (PRO):

PROs include HRQoL/functional status, symptom/symptom burden, experience with care, health-related behaviors. (*A PRO-based performance measure is not a survey instrument. Data may be collected using a survey instrument to construct a PRO measure.)*

☐ Intermediate clinical outcome (*e.g., lab value*):

☐ Process:

☐ Appropriate use measure:

☐ Structure:

☐ Composite:

1a.2 LOGIC MODEL Diagram or briefly describe the steps between the healthcare structures and processes (e.g., interventions, or services) and the patient's health outcome(s). The relationships in the diagram should be easily understood by general, non-technical audiences. Indicate the structure, process or outcome being measured.

Postoperative complications and operative mortality are important negative outcomes associated with lung cancer resection surgery. The STS lung cancer resection risk model (Fernandez et al, 2016) identifies predictors of these outcomes, including patient age, smoking status, comorbid medical conditions, and other patient characteristics, as well as operative approach and the extent of pulmonary resection. Knowledge of these predictors informs clinical decision

making by enabling physicians and patients to understand the associations between individual patient characteristics and outcomes and – with continuous feedback of performance data over time – fosters quality improvement.

Fernandez FG, Kosinski AS, Burfeind W, et al. The Society of Thoracic Surgeons lung cancer resection risk model: higher quality data and superior outcomes. Ann Thorac Surg 2016;102:370-7.

1a.3 Value and Meaningfulness: IF this measure is derived from patient report, provide evidence that the target population values the measured ***outcome, process, or structure*** and finds it meaningful. (Describe how and from whom their input was obtained.)

n/a

****RESPOND TO ONLY ONE SECTION BELOW -EITHER 1a.2, 1a.3 or 1a.4) ****

1a.2 FOR OUTCOME MEASURES including PATIENT REPORTED OUTCOMES - Provide empirical data demonstrating the relationship between the outcome (or PRO) to at least one healthcare structure, process, intervention, or service.

The STS lung cancer resection data demonstrate a significant relationship between operative approach (i.e., thoracoscopy vs. thoracotomy), postoperative complications and operative mortality. Please see Table 4 in the attachment (Fernandez et al, 2016) for empirical data related to operative approach and also for procedure type/extent of pulmonary resection.

Fernandez FG, Kosinski AS, Burfeind W, et al. The Society of Thoracic Surgeons lung cancer resection risk model: higher quality data and superior outcomes. Ann Thorac Surg 2016;102:370-7.

1a.3. SYSTEMATIC REVIEW(SR) OF THE EVIDENCE (for INTERMEDIATE OUTCOME, PROCESS, OR STRUCTURE PERFORMANCE MEASURES, INCLUDING THOSE THAT ARE INSTRUMENT-BASED) If the evidence is not based on a systematic review go to section 1a.4) If you wish to include more than one systematic review, add additional tables.

What is the source of the systematic review of the body of evidence that supports the performance measure? A systematic review is a scientific investigation that focuses on a specific question and uses explicit, prespecified scientific methods to identify, select, assess, and summarize the findings of similar but separate studies. It may include a quantitative synthesis (meta-analysis), depending on the available data. (IOM)

☐ Clinical Practice Guideline recommendation (with evidence review)

☐ US Preventive Services Task Force Recommendation

☐ Other systematic review and grading of the body of evidence (*e.g., Cochrane Collaboration, AHRQ Evidence Practice Center*)

☐ Other

Source of Systematic Review: • **Title** • **Author** • **Date** • **Citation, including page number** • **URL**	
Quote the guideline or recommendation verbatim about the process, structure or intermediate outcome being measured. If not a guideline, summarize the conclusions from the SR.	
Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	
Provide all other grades and definitions from the evidence grading system	
Grade assigned to the **recommendation** with definition of the grade	
Provide all other grades and definitions from the recommendation grading system	

Body of evidence: • Quantity – how many studies? • Quality – what type of studies?	
Estimates of benefit and consistency across studies	
What harms were identified?	
Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	

1a.4 OTHER SOURCE OF EVIDENCE

If source of evidence is NOT from a clinical practice guideline, USPSTF, or systematic review, please describe the evidence on which you are basing the performance measure.

1a.4.1 Briefly SYNTHESIZE the evidence that supports the measure. A list of references without a summary is not acceptable.

1a.4.2 What process was used to identify the evidence?

1a.4.3. Provide the citation(s) for the evidence.

1b. Performance Gap

Demonstration of quality problems and opportunity for improvement, i.e., data demonstrating:

- considerable variation, or overall less-than-optimal performance, in the quality of care across providers; and/or
- Disparities in care across population groups.

1b.1. Briefly explain the rationale for this measure *(e.g., how the measure will improve the quality of care, the benefits or improvements in quality envisioned by use of this measure)*

If a COMPOSITE (e.g., combination of component measure scores, all-or-none, any-or-none), SKIP this question and answer the composite questions.

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 enable them to compare risk-adjusted outcomes for quality improvement purposes.

1b.2. Provide performance scores on the measure as specified (current and over time) at the specified level of analysis. *(This is required for maintenance of endorsement. Include mean, std dev, min, max, interquartile range, scores by decile. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities include.) This information also will be used to address the sub-criterion on improvement (4b1) under Usability and Use.*

The endpoint of mortality or major morbidity occurred in 9.5% of eligible patients. There is no overlap in credible intervals for hospital-specific SIR between some of the best performing sites (3.5%; 8 of 231 sites with upper limit below 1) and worst performing sites (6.9%; 16 of 231 sites with lower limit above 1), indicating that this model provides meaningful discrimination between best and worst performers.

Dates: January 1, 2012 through December 31, 2014

Data/Sample: 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]).

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

Minimum	**0.47**
1st quartile	0.90
Median	1.00
Mean	1.05
3rd quartile	1.22
Maximum	2.37

1b.3. If no or limited performance data on the measure as specified is reported in 1b2, then provide a summary of data from the literature that indicates opportunity for improvement or overall less than optimal performance on the specific focus of measurement.

n/a (see data reported in 1b2)

1b.4. Provide disparities data from the measure as specified (current and over time) by population group, e.g., by race/ethnicity, gender, age, insurance status, socioeconomic status, and/or disability. *(This is required for maintenance of endorsement. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included.) For measures that show high levels of performance, i.e., "topped out", disparities data may demonstrate an opportunity for improvement/gap in care for certain sub-populations. This information also will be used to address the sub-criterion on improvement (4b1) under Usability and Use.*

Data/Sample: The population included 27,844 records from 231 hospitals.

Dates: January 1, 2012 through December 31, 2014

Race: White 24,099; Black 2,369; Other 1,217

Incidence of mortality or major morbidity endpoints:

White: 9.8%, 95% CI [9.4,% 10.1%]

Black: 8.9%, 95% CI [7.8%, 10.1%]

Other: 6.9%, 95% CI [5.6, 8.5%]

1b.5. If no or limited data on disparities from the measure as specified is reported in 1b.4, then provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement. Include citations. Not necessary if performance data provided in 1b.4

n/a (see data reported in 1b4)

2. Reliability and 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. ***Measures must be judged to meet the sub criteria for both reliability and validity to pass this criterion and be evaluated against the remaining criteria.***

2a.1. Specifications The measure is well defined and precisely specified so it can be implemented consistently within and across organizations and allows for comparability. eMeasures should be specified in the Health Quality Measures Format (HQMF) and the Quality Data Model (QDM).

De.5. Subject/Topic Area *(check all the areas that apply):*

Cancer, Cancer : Lung, Esophageal, Surgery, Surgery : Thoracic Surgery

De.6. Non-Condition Specific*(check all the areas that apply):*

Safety, Safety : Complications

De.7. Target Population Category *(Check all the populations for which the measure is specified and tested if any):*

Elderly

S.1. Measure-specific Web Page *(Provide a URL link to a web page specific for this measure that contains current detailed specifications including code lists, risk model details, and supplemental materials. Do not enter a URL linking to a home page or to general information.)*

http://www.sts.org/sites/default/files/documents/STSThoracicDataSpecsV2_3.pdf

S.2a. If this is an eMeasure, HQMF specifications must be attached. Attach the zipped output from the eMeasure authoring tool (MAT) - if the MAT was not used, contact staff. (Use the specification fields in this online form for the plain-language description of the specifications)

This is not an eMeasure **Attachment:**

S.2b. Data Dictionary, Code Table, or Value Sets *(and risk model codes and coefficients when applicable) must be attached. (Excel or csv file in the suggested format preferred - if not, contact staff)*

Attachment **Attachment:** STSThoracicDataSpecsV2_3.pdf

S.2c. Is this an instrument-based measure (i.e., data collected via instruments, surveys, tools, questionnaires, scales, etc.)? Attach copy of instrument if available.

No, this is not an instrument-based measure **Attachment:**

S.2d. Is this an instrument-based measure (i.e., data collected via instruments, surveys, tools, questionnaires, scales, etc.)? Attach copy of instrument if available.

Not an instrument-based measure

S.3.1. For maintenance of endorsement: Are there changes to the specifications since the last updates/submission. If yes, update the specifications for S1-2 and S4-22 and explain reasons for the changes in S3.2.

Yes

S.3.2. For maintenance of endorsement, please briefly describe any important changes to the measure specifications since last measure update and explain the reasons.

Among postoperative complications included in the numerator statement, "bleeding requiring reoperation" was replaced by "unexpected return to the operating room." Bleeding is only one of many possible reasons for a reoperation; other reasons may include prolonged air leak and chylothorax. STS General Thoracic surgeon leaders felt that the new, expanded definition of reoperation ("unexpected return to the operating room") better reflects the scope of this category of postoperative complications.

S.4. Numerator Statement *(Brief, narrative description of the measure focus or what is being measured about the target population, i.e., cases from the target population with the target process, condition, event, or outcome) DO NOT include the rationale for the measure.*

IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm (S.14).

Number of patients greater than or equal to 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, unexpected return to the operating room, myocardial infarction or operative mortality (death during the index hospitalization, regardless of timing, or within 30 days, regardless of location).

S.5. 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, time period for data collection, specific data collection*

items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)

IF an OUTCOME MEASURE, describe how the observed outcome is identified/counted. Calculation of the risk-adjusted outcome should be described in the calculation algorithm (S.14).

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. Acute 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"

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 2.3-

http://www.sts.org/sites/default/files/documents/STSThoracicDCF_V2_3_MajorProc_Annotated.pdf

S.6. Denominator Statement *(Brief, narrative description of the target population being measured)*

Number of patients greater than or equal to 18 years of age undergoing elective lung resection (Open or VATS wedge resection, segmentectomy, lobectomy, bilobectomy, sleeve lobectomy, pneumonectomy) for lung cancer

S.7. Denominator Details *(All information required to identify and calculate the target population/denominator such as definitions, time period for data collection, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b.)*

IF an OUTCOME MEASURE, describe how the target population is identified. Calculation of the risk-adjusted outcome should be described in the calculation algorithm (S.14).

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 (32666)

Thoracoscopy with removal of a single lung segment (segmentectomy) (32669)

Thoracoscopy with removal of two lobes (bilobectomy) (32670)

Thoracoscopy with removal of lung, pneumonectomy (32671)

Thoracotomy with therapeutic wedge resection (eg mass nodule) initial (32505)

Removal of lung, total pneumonectomy; (32440)

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)

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

S.8. Denominator Exclusions *(Brief narrative description of exclusions from the target population)*

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. Furthermore, patients with missing age, sex, discharge mortality status, and predicted forced expiratory volume in 1 second were also excluded.

S.9. Denominator Exclusion Details *(All information required to identify and calculate exclusions from the denominator such as definitions, time period for data collection, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b.)*

Cases removed from calculations if any of following fields are checked on the data collection form:

Removal of lung, sleeve (carinal) pneumonectomy (32442)

Removal of lung, total pneumonectomy; extrapleural (32445)

Removal of lung, completion pneumonectomy (32488)

OR if either of the following fields are checked:

Carcinoid tumor of bronchus and lung; benign, typical (209.61., D34.090)

Lung tumor, benign (212.3, D14.30)

OR if Emergent, Urgent, or Palliative is checked under "Status of Operation"

Only general thoracic procedures coded as primary lung or primary esophageal cancer are included in measure calculations, so occult carcinoma is effectively excluded.

S.10. Stratification Information *(Provide all information required to stratify the measure results, if necessary, including the stratification variables, definitions, specific data collection items/responses, code/value sets, and the risk-model covariates and coefficients for the clinically-adjusted version of the measure when appropriate – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format with at S.2b.)*

n/a

S.11. Risk Adjustment Type (Select type. Provide specifications for risk stratification in measure testing attachment)

Statistical risk model

If other:

S.12. Type of score:

Rate/proportion

If other:

S.13. 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

S.14. Calculation Algorithm/Measure Logic (*Diagram or 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; time period for data, aggregating data; risk adjustment; etc.)*

Target population is patients undergoing elective lung resection for lung cancer. Emergency procedures were excluded. Outcome is operative mortality (death during the index hospitalization, regardless of timing, or within 30 days, regardless of location) or occurrence of any of the following postoperative complications: reintubation, need for tracheostomy, initial ventilator support > 48 hours, ARDS, pneumonia, pulmonary embolus, bronchopleural fistula, unexpected return to the operating room, or myocardial infarction. Analysis considered 27,844 patients with procedures between 01/01/2012 and 12/31/2014 (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, please review the risk model in the attachment. (Fernandez, et. al. 2016.)

S.15. Sampling *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*

IF an instrument-based performance measure (e.g., PRO-PM), identify whether (and how) proxy responses are allowed.

n/a

S.16. Survey/Patient-reported data *(If measure is based on a survey or instrument, provide instructions for data collection and guidance on minimum response rate.)*

Specify calculation of response rates to be reported with performance measure results.

n/a

S.17. Data Source *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*

If other, please describe in S.18.

Other, Registry Data

S.18. Data Source or Collection Instrument *(Identify the specific data source/data collection instrument (e.g. name of database, clinical registry, collection instrument, etc., and describe how data are collected.)*

IF instrument-based, identify the specific instrument(s) and standard methods, modes, and languages of administration.

STS General Thoracic Surgery Database, Version 2.3

S.19. Data Source or Collection Instrument *(available at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*

Available at measure-specific web page URL identified in S.1

S.20. Level of Analysis *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*

Facility

S.21. Care Setting *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*

Inpatient/Hospital

If other:

S.22. COMPOSITE Performance Measure - Additional Specifications *(Use this section as needed for aggregation and weighting rules, or calculation of individual performance measures if not individually endorsed.)*

2. Validity – See attached Measure Testing Submission Form

2.1 For maintenance of endorsement

Reliability testing: If testing of reliability of the measure score was not presented in prior submission(s), has reliability testing of the measure score been conducted? If yes, please provide results in the Testing attachment. Please use the most current version of the testing attachment (v7.1). Include information on all testing conducted (prior testing as well as any new testing); use red font to indicate updated testing.

Yes

2.2 For maintenance of endorsement

Has additional empirical validity testing of the measure score been conducted? If yes, please provide results in the Testing attachment. Please use the most current version of the testing attachment (v7.1). Include information on all testing conducted (prior testing as well as any new testing); use red font to indicate updated testing.

Yes

2.3 For maintenance of endorsement

Risk adjustment: For outcome, resource use, cost, and some process measures, risk-adjustment that includes social risk factors is not prohibited at present. Please update sections 1.8, 2a2, 2b1,2b4.3 and 2b5 in the Testing attachment and S.140 and S.11 in the online submission form. NOTE: These sections must be updated even if social risk factors are not included in the risk-adjustment strategy. You MUST use the most current version of the Testing Attachment (v7.1) -- older versions of the form will not have all required questions.

Yes - Updated information is included

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.
- 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] 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] 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]

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]

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; [14,15] 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] **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:	☐ other:

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 <u>all</u> 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:	☐ other:

1.5. How many and which <u>measured entities</u> 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 <u>patients</u> 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].

Variable	**Values**
Total	27,844 (100)
Age, years	67.2 ± 10.1
Male	12,647 (45.4)
Race	
White	24,099 (87.0)
Black	2,369 (8.6)
Other	1,217 (4.4)
Body mass index, kg/m^{2a}	27.6 ± 6.2
Coronary artery disease	6,196 (22.3)
Diabetes mellitus	5,158 (18.5)
Renal dysfunction	504 (1.8)
Induction chemotherapy or radiation	1,801 (6.5)

Variable	Values
Cigarette smoking	
Never	3,895 (14.0)
Past (stopped more than 1 month)	17,368 (62.4)
Current	6,581 (23.6)
Steroids	965 (3.5)
Minimally invasive	17,153 (61.6)
Thoracotomy	10,691 (38.4)
Primary procedure	
Wedge resection	3,815 (13.7)
Segmentectomy	1,685 (6.1)
Lobectomy	19,836 (71.2)
Sleeve lobectomy	412 (1.5)
Bilobectomy	980 (3.5)
Pneumonectomy	1,116 (4.0)

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 (ρ^2) between the set of participant-specific estimates

$\hat{\theta}_1, \ldots, \hat{\theta}_N$ and the corresponding unknown true values, $\theta 1, \ldots, \theta_N$, that is:

$$\rho^2 = \frac{\sum_{j=1}^{N}(\hat{\theta}_j - \frac{1}{N}\sum_{h=1}^{N}\hat{\theta}_h)(\theta_j - \frac{1}{N}\sum_{h=1}^{N}\theta_h)}{\sum_{j=1}^{N}(\hat{\theta}_j - \frac{1}{N}\sum_{h=1}^{N}\hat{\theta}_h)^2 \sum_{j=1}^{N}(\theta_j - \frac{1}{N}\sum_{h=1}^{N}\theta_h)^2}$$

The quantity ρ^2 was estimated by its posterior mean, namely,

$$\hat{\rho}^2 = \frac{1}{5000}\sum_{l=1}^{5000}\rho_{(l)}^2$$

where

$$\rho_{(l)}^2 = \frac{\sum_{j=1}^{N}(\hat{\theta}_j - \frac{1}{N}\sum_{h=1}^{N}\hat{\theta}_h)(\theta_j^{(l)} - \frac{1}{N}\sum_{h=1}^{N}\theta_h^{(l)})}{\sum_{j=1}^{N}(\hat{\theta}_j - \frac{1}{N}\sum_{h=1}^{N}\hat{\theta}_h)^2 \sum_{j=1}^{N}(\theta_j^{(l)} - \frac{1}{N}\sum_{h=1}^{N}\theta_h^{(l)})^2}$$

$$\theta_j \hat{\theta}_j = \sum_{l=1}^{5000}\theta_j^{(l)}/5000 \quad \theta_h^{(l)}$$

with θ_j denoting the value of $\rho_{(l)}^2$ on the *l*-th MCMC sample $\hat{\theta}_j = \sum_{l=1}^{5000}\theta_j^{(l)}/5000$ denoting the posterior mean of θ_j. A 95% credible interval for ρ^2 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.

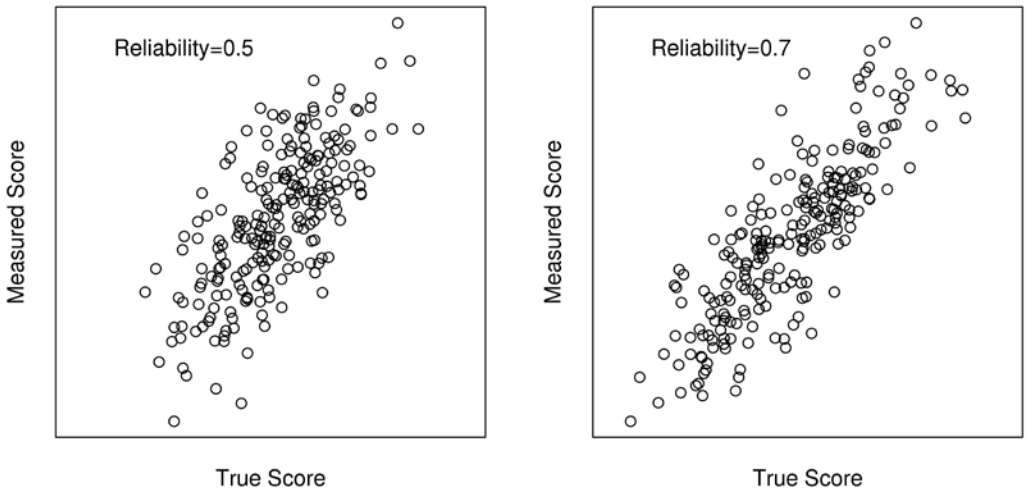
Given the timeframe of the data used for reliability testing for this measure (01/01/2012 – 12/31/2014), the revised postoperative complication data element "unexpected return to the operating room" was included in the analysis.

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 Measured Score$_i$=True Score$_i$ + e_i where $i = 1,2, \ldots ,231$ indicates the 231 participants and where True Score$_i$ and e_i 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%

CATEGORY	FIELD_NAME	NUM	DEN	Agreement Rate
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

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.**

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

☐ No risk adjustment or stratification

☒ Statistical risk model with _risk factors

☐ Stratification by _risk categories

☐ Other,

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.

Variable	Composite Model (Mortality or Major Morbidity) OR (95% CI)	*p* Value
Age, 10-year increase	1.14 (1.08-1.90)	<0.001
Male	1.41 (1.29-1.53)	<0.001
Body mass index, kg/m^2		<0.001
≥18.5 to <25	1.00	
≥6.0 to <18.5	1.35 (1.09-1.66)	
≥25.0 to <30.0	0.83 (0.75-0.92)	
≥30.0 to <35.0	0.72 (0.63-0.82)	
≥35.0 to ≤99.9	0.83 (0.71-0.97)	
Hypertension	1.06 (0.96-1.16)	0.25
Steroids	1.33 (1.09-1.62)	0.005
Congestive heart failure	1.19 (0.97-1.46)	0.10
Coronary artery disease	1.14 (1.03-1.26)	0.011
Peripheral vascular disease	1.43 (1.26-1.63))	<0.001
Reoperation	1.32 (1.13-1.54)	<0.001
Cerebrovascular disease	1.11 (0.97-1.28)	0.14
Diabetes mellitus	1.01 (0.91-1.13)	0.84
% FEV , 10% decrease	1.12 (1.10-1.15)	<0.001
Induction therapy	1.20 (1.03-1.39)	0.022
Renal dysfunction	1.11 (0.84-1.46)	0.47
Cigarette smoking		<0.001
Never	1.00	
Past smoker	1.23 (1.05-1.44)	
Current smoker	1.64 (1.38-1.94)	
Zubrod score		<0.001
0	1.00	
1	1.16 (1.06-1.28)	
2-5	1.60 (1.32-1.95)	
ASA		<0.001
1 or 2	1.00	
3	1.27 (1.09-1.47)	
4 or 5	1.76 (1.45-2.13)	
Approach		<0.001
Minimally invasive	1.00	
Thoracotomy	1.51 (1.37-1.66)	
Pathologic stage		0.25
I	1.00	
II	1.05 (0.95-1.17)	
III	1.14 (1.00-1.30)	
IV	1.04 (0.75-1.42)	
Procedure		<0.001
Wedge	1.00	
Segmentectomy	1.24 (0.97-1.57)	
Lobectomy	1.93 (1.65-2.26)	
Sleeve	1.96 (1.39-2.77)	
Bilobectomy	2.91 (2.29-3.70)	
Pneumonectomy	2.83 (2.24-3.58)	

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

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

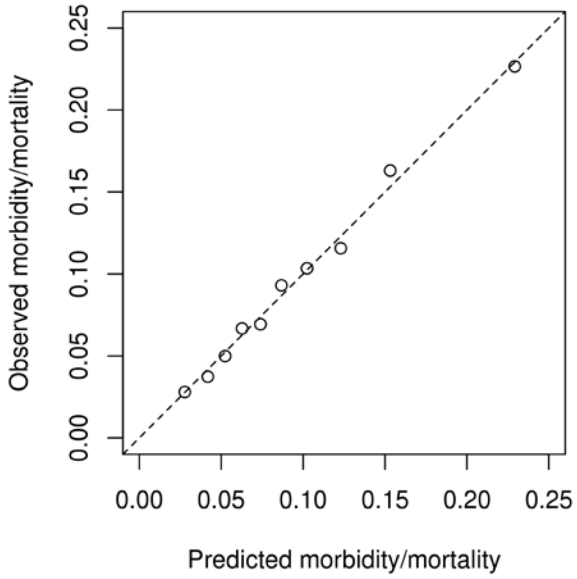
The C-statistics is 0.68.

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.

3. Feasibility

Extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

3a. Byproduct of Care Processes

For clinical measures, the required data elements are routinely generated and used during care delivery (e.g., blood pressure, lab test, diagnosis, medication order).

3a.1. Data Elements Generated as Byproduct of Care Processes.

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)

If other:

3b. Electronic Sources

The required data elements are available in electronic health records or other electronic sources. If the required data are not in electronic health records or existing electronic sources, a credible, near-term path to electronic collection is specified.

3b.1. To what extent are the specified data elements available electronically in defined fields (*i.e., data elements that are needed to compute the performance measure score are in defined, computer-readable fields*) Update this field for **maintenance of endorsement**.

ALL data elements are in defined fields in a combination of electronic sources

3b.2. If ALL the data elements needed to compute the performance measure score are not from electronic sources, specify a credible, near-term path to electronic capture, OR provide a rationale for using other than electronic sources. For **maintenance of endorsement**, if this measure is not an eMeasure (eCQM), please describe any efforts to develop an eMeasure (eCQM).

n/a

3b.3. If this is an eMeasure, provide a summary of the feasibility assessment in an attached file or make available at a measure-specific URL. Please also complete and attach the NQF Feasibility Score Card.

Attachment:

3c. Data Collection Strategy

Demonstration that the data collection strategy (e.g., source, timing, frequency, sampling, patient confidentiality, costs associated with fees/licensing of proprietary measures) can be implemented (e.g., already in operational use, or testing demonstrates that it is ready to put into operational use). For eMeasures, a feasibility assessment addresses the data elements and measure logic and demonstrates the eMeasure can be implemented or feasibility concerns can be adequately addressed.

3c.1. Required for maintenance of endorsement. Describe difficulties (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.

IF instrument-based, consider implications for both individuals providing data (patients, service recipients, respondents) and those whose performance is being measured.

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 $550 or $700, depending on whether the participant is an STS member or not. STS membership thus provides surgeons with a 21% discount on the non-member database participation fee.

3c.2. Describe any fees, licensing, or other requirements to use any aspect of the measure as specified *(e.g., value/code set, risk model, programming code, algorithm)*.

See 3c.1

4. Usability and Use

Extent to which potential audiences (e.g., consumers, purchasers, providers, policy makers) are using or could use performance results for both accountability and performance improvement to achieve the goal of high-quality, efficient healthcare for individuals or populations.

4a. Accountability and Transparency

Performance results are used in at least one accountability application within three years after initial endorsement and are publicly reported within six years after initial endorsement (or the data on performance results are available). If not in use at the time of initial endorsement, then a credible plan for implementation within the specified timeframes is provided.

4.1. Current and Planned Use

NQF-endorsed measures are expected to be used in at least one accountability application within 3 years and publicly reported within 6 years of initial endorsement in addition to performance improvement.

Specific Plan for Use	Current Use (for current use provide URL)
Public Reporting	Quality Improvement (external benchmarking to organizations) STS General Thoracic Surgery Database http://publicreporting.sts.org/gtsd Quality Improvement (Internal to the specific organization) STS General Thoracic Surgery Database http://publicreporting.sts.org/gtsd

4a1.1 For each CURRENT use, checked above (update for maintenance of endorsement), provide:

- Name of program and sponsor
- Purpose
- Geographic area and number and percentage of accountable entities and patients included
- Level of measurement and setting

See 4a1.2

4a1.2. If not currently publicly reported OR used in at least one other accountability application (e.g., payment program, certification, licensing) what are the reasons? (*e.g., Do policies or actions of the developer/steward or accountable entities restrict access to performance results or impede implementation?*)

STS is actively promoting public reporting of the STS adult cardiac, congenital heart, and general thoracic surgery performance measures. This is consistent with the explicitly stated STS philosophy that "As a national leader in health care transparency and accountability, The Society of Thoracic Surgeons believes that the public has a right to know the quality of surgical outcomes." (http://www.sts.org/registries-research-center/sts-public-reporting) In our efforts to operationalize public reporting, the STS Public Reporting Task Force has and will continue to develop public report cards that are consumer centric. Public reporting remains a top priority for the Society, and STS is striving for even stronger involvement among Database participants.

Currently, more than 650 Adult Cardiac Surgery Database (ACSD) participants voluntarily consent to be a part of the STS Public Reporting and more than 550 ACSD participants have consented to report publicly via the Consumer Reports public reporting initiative. Additionally, more than 100 Congenital Heart Surgery Database (CHSD) participants are currently enrolled in STS Public Reporting.

As of July 2017, General Thoracic Surgery Database (GTSD) participants were included in the Public Reporting initiative and more than 250 participants currently consent to report outcomes publicly on the STS website. This includes discharge mortality rate and median postoperative length of stay for lobectomy procedures for lung cancer, including scores and star ratings for the Lobectomy for Lung Cancer Composite Measure in addition to its domains of 1) absence

of mortality, and 2) absence of major complication. Participant outcomes are published alongside GTSD overall outcomes and National Inpatient Sample (NIS) outcomes.
-ACSD public reporting online may be found here: http://publicreporting.sts.org/acsd
-CHSD public reporting online may be found here: http://publicreporting.sts.org/chsd
-GHSD public reporting online may be found here: http://publicreporting.sts.org/gtsd

4a1.3. If not currently publicly reported OR used in at least one other accountability application, provide a credible plan for implementation within the expected timeframes -- any accountability application within 3 years and publicly reported within 6 years of initial endorsement. (*Credible plan includes the specific program, purpose, intended audience, and timeline for implementing the measure within the specified timeframes. A plan for accountability applications addresses mechanisms for data aggregation and reporting.*)

See 4a1.2

4a2.1.1. Describe how performance results, data, and assistance with interpretation have been provided to those being measured or other users during development or implementation.

How many and which types of measured entities and/or others were included? If only a sample of measured entities were included, describe the full population and how the sample was selected.

STS's combined mortality and morbidity model for pulmonary resection for lung cancer is important and appropriate for public reporting for the following reasons:

1.) lung cancer resection is the most common category of surgical procedures that a thoracic surgeon performs;

2.) these procedures are therefore useful and appropriate to use as a benchmark for performance by general thoracic surgery programs. By providing surgeons and teams with risk-adjusted results, they can identify how they are performing compared with other programs in the STS General Thoracic Database, which generally includes the top thoracic programs in the nation. This will assist them in focusing performance improvement efforts. Also, when publicly reported, the outcomes for these common procedures provide patients and their families with comparative performance information to aid in selection of a provider;

3.) major morbidity is relatively common after lung resection; however, although mortality is rare, it should be captured as well in an outcome measure, thereby identifying ALL adverse events after lung resection;

4.) this measure is reported in an easy to understand format which summarizes the results of all participants who were included in the analysis. The participant's score is illustrated graphically in relation to the 25th, 50th and 75th percentiles of the distribution across participants, and is accompanied by the 95% Bayesian credible interval. 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 compare their patients´ outcomes with national benchmarks and to initiate QI efforts as needed.

4a2.1.2. Describe the process(es) involved, including when/how often results were provided, what data were provided, what educational/explanatory efforts were made, etc.

See 4a2.1.1

4a2.2.1. Summarize the feedback on measure performance and implementation from the measured entities and others described in 4d.1.

Describe how feedback was obtained.

The general thoracic surgeons from across the U.S. who comprise the STS General Thoracic Surgery Task Force meet periodically to discuss the participant reports and to consider potential enhancements to the GTSD. Additions/clarifications to the data collection form and to the content/format of the participant reports are discussed and implemented as appropriate.

Most recently, STS surgeon members have expressed interest in real-time, online data updates, which has led to the development of dashboard-type reporting on STS.org. The general thoracic dashboard is scheduled for launch in 2018.

Also, general thoracic public reporting was initiated in the summer of 2017 (http://publicreporting.sts.org/gtsd), making star ratings for consenting participant groups available to participants as well as the public.

4a2.2.2. Summarize the feedback obtained from those being measured.

See 4a2.2.1

4a2.2.3. Summarize the feedback obtained from other users

Given the very recent launch of general thoracic public reporting, the STS has not yet received sufficient feedback from non-participants to be able to assess the impact of the public reporting initiative.

4a2.3. Describe how the feedback described in 4a2.2.1 has been considered when developing or revising the measure specifications or implementation, including whether the measure was modified and why or why not.

See Specifications section, S.3.2, regarding modification in postoperative complications included in numerator since most recent NQF review of this measure.

Improvement

Progress toward achieving the goal of high-quality, efficient healthcare for individuals or populations is demonstrated. If not in use for performance improvement at the time of initial endorsement, then a credible rationale describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.

4b1. Refer to data provided in 1b but do not repeat here. Discuss any progress on improvement (trends in performance results, number and percentage of people receiving high-quality healthcare; Geographic area and number and percentage of accountable entities and patients included.)

If no improvement was demonstrated, what are the reasons? If not in use for performance improvement at the time of initial endorsement, provide a credible rationale that describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.

Operative mortality in the STS General Thoracic Surgery Database has decreased from 2.2% in the years 2002 to 2008 to 1.4% from 2012 to 2014. These data represent the highest quality lung cancer surgery in the United States. It is important to recognize that a large proportion of the general thoracic surgery in the US is not performed by general thoracic surgeons certified by the American Board of Thoracic Surgery. Results by STS General Thoracic Database participants, who are almost all ABTS certified, are generally superior to those of surgeons performing these procedures who do not participate in the GTSD, and who are often not ABTS certified.

Kozower and colleagues (Ann Thorac Surg 2010) have previously demonstrated that compared with the Nationwide Inpatient Sample database, from 2002 to 2008, patients in the GTSD had lower unadjusted discharge mortality rates, median length of stay, and pulmonary complication rates for lobectomy.

The major morbidity rate has increased from 8.6% to 9.1% during the same time. A potential explanation for this observation is more complete coding of complications by data abstractors as the result of education efforts from STS, as well as inclusion of unexpected return to the operating room for any reason instead of only for bleeding.

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.

Kozower BD, Sheng S, O'Brien SM, et al. STS database risk models: predictors of mortality and major morbidity for lung cancer resection. Ann Thorac Surg 2010;90:875–83.

4b2. Unintended Consequences

The benefits of the performance measure in facilitating progress toward achieving high-quality, efficient healthcare for individuals or populations outweigh evidence of unintended negative consequences to individuals or populations (if such evidence exists).

4b2.1. Please explain any unexpected findings (positive or negative) during implementation of this measure including unintended impacts on patients.

We are not aware of any unexpected findings associated with implementation of this measure.

4b2.2. Please explain any unexpected benefits from implementation of this measure.

n/a

5. Comparison to Related or 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.

5. Relation to Other NQF-endorsed Measures

Are there related measures (conceptually, either same measure focus or target population) or competing measures (conceptually both the same measure focus and same target population)? If yes, list the NQF # and title of all related and/or competing measures.

Yes

5.1a. List of related or competing measures (selected from NQF-endorsed measures)

5.1b. If related or competing measures are not NQF endorsed please indicate measure title and steward.

This measure is related conceptually to the STS Lobectomy for Lung Cancer Composite Score measure, which we are submitting for initial NQF review in the fall 2017 Surgery endorsement cycle. The numerators for both measures include the same list of postoperative complications, but the outcomes for the Lobectomy Composite measure are grouped into two domains (operative mortality and major complications) and the measure is structured to provide general thoracic surgeons with a "star rating." Please also see 5a.2 below.

5a. Harmonization of Related Measures

The measure specifications are harmonized with related measures;

OR

The differences in specifications are justified

5a.1. If this measure conceptually addresses EITHER the same measure focus OR the same target population as NQF-endorsed measure(s):

Are the measure specifications harmonized to the extent possible?

Yes

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

Measure #1790 includes a broader range of lung resection procedures than the Lobectomy Composite, and therefore includes a larger number of cases and potentially provides performance data to more general thoracic surgeons. Of the two measures, only the Lobectomy Composite is currently publicly reported.

5b. Competing Measures

The measure is superior to competing measures (e.g., is a more valid or efficient way to measure);

OR

Multiple measures are justified.

5b.1. If this measure conceptually addresses 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

Appendix

A.1 Supplemental materials may be provided in an appendix. All supplemental materials (such as data collection instrument or methodology reports) should be organized in one file with a table of contents or bookmarks. If material pertains to a specific submission form number, that should be indicated. Requested information should be provided in the submission form and required attachments. There is no guarantee that supplemental materials will be reviewed.

Attachment Attachment: FernandezKosinskiKozower_lung_cancer_risk_model_2016.pdf

Contact Information

Co.1 Measure Steward (Intellectual Property Owner): The Society of Thoracic Surgeons

Co.2 Point of Contact: Mark, Antman, mantman@sts.org, 312-202-5856-

Co.3 Measure Developer if different from Measure Steward: The Society of Thoracic Surgeons

Co.4 Point of Contact: Mark, Antman, mantman@sts.org, 312-202-5856-

Additional Information

Ad.1 Workgroup/Expert Panel involved in measure development

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.

Measure Developer/Steward Updates and Ongoing Maintenance

Ad.2 Year the measure was first released: 2010

Ad.3 Month and Year of most recent revision: 02, 2016

Ad.4 What is your frequency for review/update of this measure? annually

Ad.5 When is the next scheduled review/update for this measure? 01, 2018

Ad.6 Copyright statement:

Ad.7 Disclaimers:

Ad.8 Additional Information/Comments: