NATIONAL QUALITY FORUM
Driving measurable health improvements together

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: Ctrl + click link to go to the link; ALT + LEFT ARROW to return

Brief Measure Information

NQF #: 3294
Corresponding Measures:
Measure Title: STS Lobectomy for Lung Cancer Composite Score
Measure Steward: The Society of Thoracic Surgeons
sp.02. Brief Description of Measure: The STS Lobectomy Composite Score comprises two domains:
1. Operative Mortality (death during the same hospitalization as surgery or within 30 days of the procedure)
2. Presence of at least one of these major complications: pneumonia, acute respiratory distress syndrome, bronchopleural fistula, pulmonary embolus, initial ventilator support greater than 48 hours, reintubation/respiratory failure, tracheostomy, myocardial infarction, or unexpected return to the operating room.
The composite score is created by a weighted combination of the above two domains resulting in a single composite score. In addition to receiving a numeric score, participants are assigned to rating categories designated by the following:
1 star: lower-than expected performance
2 stars: as-expected-performance
3 start: higher-than-expected-performance
1b.01. Developer Rationale: N/A

---

sp.12. Numerator Statement: The STS Lobectomy Composite Score comprises two domains:
1. Operative Mortality (death during the same hospitalization as surgery or within 30 days of the procedure)
2. Presence of at least one of these major complications: pneumonia, acute respiratory distress syndrome, bronchopleural fistula, pulmonary embolus, initial ventilator support greater than 48 hours, reintubation/respiratory failure, tracheostomy, myocardial infarction, or unexpected return to the operating room.
The composite score is created by a weighted combination of the above two domains resulting in a single composite score. Operative mortality and major complications were weighted inversely by their respective standard deviations across participants. This procedure is equivalent to first rescaling mortality and complications by their respective standard deviations and then assigning equal weighting to the rescaled mortality rate and rescaled complication rate. This is the same methodology used for other STS composite measures.
In addition to receiving a numeric score, participants are assigned to rating categories designated by the following:
1 star: lower-than expected performance
2 stars: as-expected-performance
3 start: higher-than-expected-performance
Patient Population: The STS GTSD was queried for all patients treated with lobectomy for lung cancer between January 1, 2014, and December 31, 2016. We excluded patients with non-elective status, occult or stage 0 tumors, American Society of Anesthesiologists class VI, and with missing data for age, sex, or discharge mortality status.
Time Window: 01/01/2014 - 12/31/2016
Model variables: Variables in the model: age, sex, year of operation, body mass index, hypertension, steroid therapy, congestive heart failure, coronary artery disease, peripheral vascular disease, reoperation, preoperative chemotherapy within 6 months, cerebrovascular disease, diabetes mellitus, renal failure, dialysis, past smoker, current smoker, forced expiratory volume in 1 second percent of predicted, Zubrod score (linear plus quadratic), American Society of Anesthesiologists class (linear plus quadratic), and pathologic stage.

sp.14. Denominator Statement: Number of patients greater than or equal to 18 years of age undergoing elective lobectomy for lung cancer
sp.16. Denominator Exclusions: Patients were excluded with non-elective status, occult or stage 0 tumors, American Society of Anesthesiologists class VI, and with missing data for age, sex, or discharge mortality status.

Measure Type: Composite
sp.29. Data Source: Other
sp.07. Level of Analysis: Facility

IF Endorsement Maintenance – Original Endorsement Date: 06/06/2018
Most Recent Endorsement Date: 6/6/2018

IF this measure is included in a composite, NQF Composite#/title: STS Lobectomy for Lung Cancer Composite Score
#3294 - STS Lobectomy for Lung Cancer Composite Score
#3294 - STS Lobectomy for Lung Cancer Composite Score
IF this measure is paired/grouped, NQF#/title:
sp.03. IF PAIRED/GROUPED, what is the reason this measure must be reported with other measures to appropriately interpret results?:

Preliminary Analysis: Maintenance of Endorsement

To maintain NQF endorsement, endorsed measures are evaluated periodically to ensure that the measure 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.

The developer provides the following description for this measure:

- This is a maintenance composite measure that assesses the operative mortality and the presence of at least one of nine major complications associated with lung cancer resection surgery, including lobectomy, the most frequently performed lung resection procedure.
- The composite score is a weighted combination of the two domains resulting in a single composite score; additionally, participants receive a star rating (1 to 3 stars) related to performance (i.e., lower-than-expected, as-expected, higher-than-expected).
- The developer provides a logic model that identifies predictors of these outcomes (e.g., age, smoking status, comorbid medical conditions, operative approach, extent of pulmonary resection) that inform

clinical decision making between physicians and patient, improve understanding of the association between individual patient characteristics and outcomes, and foster quality improvement.

Summary of prior review in 2018

- The Standing Committee agreed that the evidence supported the composite measure and that the composite score from a weighted combination of mortality and operative complications provide a more comprehensive measure of overall surgical quality.
- Overall, the Committee agreed that the quality construct and rationale for the composite are explicitly stated and logical; and the weighting and approach to the measure construction is described clearly and has been vetted by an expert panel.

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:

- The developer provided updated evidence from a 2020 study which supports minimally invasive lung cancer resection can reduce perioperative mortality and morbidity.

Question for the Committee:

- *Is there at least one thing that the provider can do to achieve a change in the measure results?*

Guidance from the Evidence Algorithm

Health Outcome or PRO (Box 1) -> Relationship between the measured health outcome and at least one healthcare action (structure, process, intervention, or service) demonstrated by empirical evidence (Box 2) -> Pass

Preliminary rating for evidence: ☒ Pass ☐ No Pass

1b. Gap in Care/Opportunity for Improvement and 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 presented the number of measured entities, number of operations, and composite score distribution estimates by percentiles and geographic region for participants with at least 30 eligible cases during two overlapping 3-year time periods (January 2017 -December 2019; January 2018 – December 2020).
    - Number of participants range from 150 to 153
    - Number of operations range from 23,292 to 24,477
    - Mean= 0.979 (Standard deviation [SD]= 0.006)- 0.98 (SD=0.005)
    - Interquartile range= 0.006-0.008
    - Minimum (0.968-0.952); Maximum (0.99-0.989)
    - Those that are in the South had a higher number of cases (n=52) compared to participants located in the Midwest (n=33), West (n=25), and Northeast (n=43).

Disparities

- The developer reported disparity data by race, ethnicity, and gender (January 2018-December 2020) that indicated that cases are:

- Higher in those that are White (n=20,507; 82.58 percent) compared to those that are Black (n=2,047; 8.24 percent), Hispanic (n=728; 2.93 percent), or Asian (n=952; 3.83 percent).
- Higher among females (n=14,098; 56.77 percent) compared to those that are male (n=10,736; 43.23 percent).
- Higher for those less than 65 years of age with commercial/HMO insurance (n=5,496; 65.25 percent) and those greater than or equal to 65 years of age with Medicare and commercial insurance without Medicaid (n=9,149; 55.75 percent).

Questions for the Committee:

- *Is there a gap in care that warrants a national performance measure?*
- *Is there opportunity for improvement based on the gap mean presented by the developer (mean=0.979)?*
- *What does the disparity data related to cases indicate (i.e., the number of operations, number of complications)?*

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

1c. Composite – Quality Construct and Rationale

Maintenance measures – same emphasis on quality construct and rationale as for new measures.

1c. Composite Quality Construct and Rationale. The quality construct and rationale should be explicitly articulated and logical; a description of how the aggregation and weighting of the components is consistent with the quality construct and rationale also should be explicitly articulated and logical.

- The measure construct is a combination of two or more individual performance measure scores (operative mortality outcome and the risk adjusted occurrence of any of nine major complications) combined into one score.
    - Operative mortality is described as death during the same hospitalization as surgery or within 30 days of the procedure.
    - Complications include:
        - Pneumonia
        - Acute respiratory distress syndrome
        - Bronchopleural fistula
        - Pulmonary embolus
        - Initial ventilator support greater than 48 hours
        - Reintubation/respiratory failure
        - Tracheostomy
        - Myocardial infarction
        - Unexpected return to the operating room
- Participants are scored for each domain (mortality and complication), and an overall composite score which is created by a weighted combination of the two domains. Participants are also assigned a rating designated by one to three stars:
    - 1 star: lower-than expected performance
    - 2 stars: as-expected performance
    - 3 stars: higher than expected performance
- The developer reports that since mortality rates for thoracic surgery have declined, it can be challenging to differentiate performance based on mortality alone since it fails to consider that not all operative survivors received equal quality care. Therefore, a composite score from a weighted combination of mortality and operative complications provides a more comprehensive measure of overall surgical quality.

- Operative mortality is weighted approximately four times that of a major complication in the composite.

Questions for the Committee:

- *Are the quality construct and a rationale for the composite explicitly stated and logical?*
- *Is the method for aggregation and weighting of the components explicitly stated and logical?*

Preliminary rating for composite quality construct and rationale: ☐ High ☒ Moderate ☐ Low ☐ Insufficient

Committee Pre-evaluation Comments:

1a. Evidence

- Moderate level evidence
- Composite measure at the facility level. The developer provided updated evidence from a 2020 study that supported the measure - minimally invasive lung cancer resection can reduce perioperative mortality and morbidity.
- Yes, evidence was provided to support the main
- New evidence that minimally invasive surgery has lower risk
- Updated evidence supports the measure and developer shared the new evidence.

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

- A performance gap exists
- The developer provided data that cases are higher in White 82.58%, compared to Black 8.24%, Hispanic 2.93%, or Asian 3.83%. There was also a gap of females 56.77% and males 43.23% and higher for those less than 65 years of age with commercial insurance 65.25%. There is room for improvement.
- Yes, it was and indicated a degree of disparity
- Gap demonstrated and also disparities
- While there is some improvement with a decline in mortality rates there is still the need to improve performance especially in persons of color and underinsured.

1c. Composite – Quality Construct and Rationale

- Moderate rating for construct and rationale
- The composite measure is operative mortality outcome and the risk adjusted occurrence of any of the nine major complications included in the measure specifications. Scoring is in mortality and also in complications. Both the mortality and complications are needed to determine and assign a quality score. Mortality is weighted approximately four times that of a major complication (the nine in the measure specs). The rationale for the composite measure is explicitly stated and logical.
- Yes
- The quality construct is clearly and logically stated.
- yes

Criteria 2: Scientific Acceptability of Measure Properties

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

Evaluators: NQF Staff

2a. Reliability: Specifications and Testing

For maintenance measures – no change in emphasis – specifications should be evaluated the same as with new measures.

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 – less emphasis if no new testing data provided.

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.

Specifications:

- Measure specifications have not changed since last review.
- Measure specifications are clear and precise.
- Measure specifications for the composite performance measure also include component measure specifications; aggregation and weighting rules; handling of missing data; standardizing scales across component measures; required sample sizes.

Reliability Testing:

- Reliability testing conducted at the Accountable Entity Level:
    - The developer conducted a signal to noise analysis using the STS General Thoracic Surgery Database (GTSD) Version 2.3 (n=233 facilities) and noted ranges from 44.6% (95% credible internal [CrI]= 34.6%-54.1%) to 68% (95% CrI= 53.6%-79.7%).
    - Providers with at least 30, 50, and 100 cases has reliability scores of 51.7%, 56.1%, and 60.9%, respectively. Large-volume participants (at least 150 cases) has a reliability of 68.0%.
    - The developer notes that while a 0.70 reliability score shows a very close correlation between measured scores, a 0.50 reliability score indicates moderate reliability and demonstrates strong correlation between the true and measures valued of the score.

Questions for the Committee regarding reliability:

- *Do you have any concerns that the measure cannot be consistently implemented (i.e., are measure specifications adequate)?*

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

2b. Validity: Validity testing; Exclusions; Risk-Adjustment; Meaningful Differences; Comparability; Missing Data

For maintenance measures – less emphasis if no new testing data provided.

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.

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

Validity Testing

- Validity testing conducted at the Accountable Entity Level:
    - The developer conducted empirical testing of the composite measure score among three-star rating performance categories (high, average, and low) and assessed composite score stability across two consecutive overlapping reporting periods.

- Providers receiving three stars had lower observed mortality risk (0.4 percent vs. 2.9 percent) and morbidity risk (2.3 percent vs. 20.0 percent) compared to participants receiving one star.
- Composite stability was assessed among 654 participants with at least 10 eligible cases using Pearson's correlation (0.71) and Spearman's correlation (0.74).
- The developer noted an increase in data accuracy rates and a narrowing of agreement ranges, indicating greater consistency in data accuracy and a high degree of data validity.

Exclusions

- The developer noted that for the measure to consistently quantify surgical quality outcomes for patients undergoing lobectomy for lung cancer, it is necessary and clinically appropriate to exclude these cases (non-elective status, missing pathology or occult or stage 0 tumors, American Society of Anesthesiologists class VI).
- The developer noted that the overall measure exclusions were 3.2% (810 of 25,640 patient records).

Risk-Adjustment

- The developer used a statistical risk model with 20 risk factors.
- Risk-adjusted operative mortality and major complications rates were estimated using a bivariate random-effects logistic regression model.
- No social risk factors were used in the statistical risk model or for stratification.
- The developer evaluated continuous variables with respect to linearity of effect.
- The model's calibration and discrimination were assessed using Hosmer-Lemeshow statistic and C-statistic.
- The C-statistic for operative mortality and major morbidity was 0.774 and 0.666, respectively and the developer interpreted these results as the model having good calibration.
- The Hosmer and Lemeshow Goodness-of-Fit Test results for operative mortality (Chi-square=14.52, degree of freedom [df]= 8, p-value= 0.07 and major morbidity (Chi-Square= 5.54, df=8, p-value=0.07) were interpreted as having good discrimination power and suitable for controlling for differences in case-mix between centers.

Meaningful Differences

- The developer determined the degree of uncertainty surrounding a participant's composite measure estimate by calculating a 95% Bayesian credible interval (Cis) and reporting the point estimates and Cis for individual participants along with a comparison to the overall average STS composite score.
- Composite measure results were converted into star categories (n=3); 91.5% of participants (n=140) with at least 30 cases over a three-year period received two stars (performance not statistically significant from overall STS national average).
- The developer notes that the identified difference among participants is both statistically significant and clinically meaningful, and the surgeon panel and users are satisfied with the distribution of participants across performance categories.

Missing Data

- Missing values were imputed when records with missing values of model covariates were identified (except for age, gender, and pathologic stage); patient records missing age, gender, or pathologic state were excluded.

- Values were imputed utilizing the median of observed variables within a category and, for binary risk factors, missing values were considered as absence of the risk factor.
- The range of missing values was between 1% and 3.5%; the developer concluded that there was no bias because of systematic missing data.

Comparability

- The measure only uses one set of specifications for this measure.

Questions for the Committee regarding validity:

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

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

2c. Composite – Empirical Analysis

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.

- The developer calculated the operative mortality and major complication rates of 186 hospitals performing at least 30 lobectomies over a three-year period to verify the contribution of each domain to the composite construction.
- The results demonstrate a reduction in mortality and major complication rates from one-star (below average) to three-star (above average) participants.
    - Mortality: 2.1 percent (one star); 1.3 percent (two star); 1.2 percent (three star)
    - Major complications: 16.2 percent (one star); 8.4 percent (two star); 3.2 percent (three star)
- Both domains were divided by their respective standard deviations across STS participants and then added together. An expert panel assessed this weighting and determined that it was consistent with their clinical assessment of each domain's relative importance.
    - Risk-standardized mortality relative weight (0.827)
    - Risk-standardized major morbidity weight (0.173)
- A one percent point change in risk-adjusted mortality rate has the same impact as a 4.8 percentage point change in the site's risk-adjusted morbidity rate.
- The developer notes that while risk-adjusted morbidity explains more of the variation in the overall composite score, it does not dominate, and both domains contribute statistical information.

Questions for the Committee regarding composite construction:

- *Do you have any concerns regarding the composite construction approach (e.g., do the component measures fit the quality construct and add value to the overall composite? Are the aggregation and weighting rules consistent with the quality construct and rationale while achieving the related objective of simplicity to the extent possible?)?*

Preliminary rating for composite construction: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee Pre-evaluation Comments:

2a. Reliability-Specification

- No concerns
- The measure specifications have not changed since the prior review. The measure specifications are clear and precise.
- no concerns
- This can be consistently implemented
- The measure specifications are adequate.

2a2. Reliability-Testing

- No concerns
- The developer conducted reliability testing at the Accountable Entity level. A signal to noice analysis using the STS General Thoracic Surgery Database Version 2.3 in 233 facilities and noted ranges from 44.6% to 68%. Reliability of the measure of providers with 30, 50 and 100 cases had reliability scores of 51.7%, 56.1% and 60.9%. Large-volume practices with at least 150 patients ranged 53.6% - 79.7%. The data for the measure is available and it should be able to be measured consistently.
- no concerns
- No concerns
- No there was no change.

2b. Validity-Testing

- No concerns
- The developer conducted empirical testing of the composite measure score using three rating categories, high, average and low across two consecutive reporting periods. Composite stability was assessed using Pearson's correlation and Spearman's correlation. The developer noted an increase in data accuracy rates and a narrowing of agreement ranges, indicating greater consistency in data accuracy and a high degree of data validity.
- no concerns
- No concern
- To Note: No SDOH risk factors where used and no risk adjustments. Otherwise straight forward for a facility quality measure.

2b2-2b3. Potential threats to validity

- Exclusions are in-line. Provided a statistical risk model but no SDOH were used.
- No change from previous
- yes
- Exclusions were consistent with the evidence. The risk adjustment used a risk model with 20 risk factors. No social risk factors were included. The model's calibration and discrimination were assessed using Hosmer-Lemesow statistic and C-statistic. The Goodness-of-Fit test results were interpreted by the developer as having good discrimination ower and suitable for controlling for differences in case-mix between centers.
- No concerns

2b4-2b7. Potential threats to validity

- Used only one specification set therefore no comparability of performance scores. No bias from the systematic missing data as tested by the developer.

- No threats
- no concerns
- The developer notes that the identified difference among participants is both statistically significant and clinically meaningful, and the surgeon panel and users are satisfied with the distribution of participants across performance categories. There was only one set of measure specifications. Missing data fields were addressed in the measure specifications. Missing values were imputed when records with missing values of model covariates were identified (except for age, gender, and pathologic stage); patient records missing age, gender, or pathologic state were excluded.
- No concerns

2c. Composite – Empirical Analysis

- No concerns
- The component measures demonstrate a reduction in mortality and major complication rates from one-star (below average) to three-star (above average) participants. Both domains were divided by their respective standard deviations across STS participants and then added together. An expert panel assessed this weighting and determined that it was consistent with their clinical assessment of each domain's relative importance.
- yes, it does
- Yes
- Yes the composite adds value and rationale.

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.

- Data elements need to compute the measure score can be generated in the following ways: generated or collected by and used by healthcare personnel during the provision of care; coded by someone other than person obtaining original information; or abstracted from a record by someone other than person obtaining original information.
- All data elements needed to compute the performance measure score are in defined fields in a combination of electronic sources.
- STS GTSD participants submit all data elements in electronic format using a standard set of data specifications.
- The developer notes that the data element variables in this composite have been standard in the STS GTSD for at least three years, with some variables for 20 years.
- Two snapshot periods (i.e., data harvest) occur annually; analyses of these results and near real-time reports are available to participants via their database platform.
- The developer notes two harvest delays due to the COVID-19 pandemic and the 2020 STS National Database warehouse transition which have been addressed.
- On-staff managers or third party abstraction companies collect these data.
- Participants pay an annual fee per surgeon whether they are a STS member or not.

Questions for the Committee:

- *Are the required data elements routinely generated and used during care delivery?*

- *Are the required data elements available in electronic form, e.g., I or other electronic sources?*
- *Is the data collection strategy ready to be put into operational use?*
- *For data elements assessed to have feasibility issues, does the developer present a credible, near-term path to electronic collection?*

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

Committee Pre-evaluation Comments:

3. Feasibility

- Moderate feasibility
- All data elements needed to compute the performance measure score are in defined fields in a combination of electronic sources. The developer notes that the data element variables in this composite have been standard in the STS GTSD for at least three years, with some variables for 20 years.
- no concerns
- all routinely generated and available electronically. No concerns
- Those who chose to participate in the STS GTSD submit a standard set of electronic generated data elements. The STS GTSD participants pay a fee per surgeon.

Criterion 4: Use and Usability

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 evaluates 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	
Planned use in an accountability program?	☐ Yes	☐ No	☒ NA	

Accountability program details

- The developer publishes measure results of consenting STS National Database participants on its website. The STS public reporting website is updated with new data once a year.
- As of March 2022, approximately 47 percent of STS General Thoracic Surgery Database (GTSD) participants were enrolled in public reporting and receive participant-level results on the following:

discharge mortality; median postoperative length of stay for lobectomy procedures for lung cancer; and STS GTSD and National Inpatient Sample (NIS) benchmarks.

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

- Participant scores are illustrated graphically in relation to the 25th, 50th, and 75th percentiles of the distribution across participants, which can easily show surgeons how they perform compared to their peers.
- The performance report includes separate domain scores, the overall composite score, and a detailed overview of the statistical calculations, endpoints, and result interpretations.
- STS General Thoracic Surgery Task Force meets periodically to discuss participant reports, potential enhancements, updates to the data collection form, and the content or format of the participant reports.
- It is unclear from the developer submission what feedback was obtained from those being measured and other users and how the feedback is considered when developing or revising the measure specifications.

***Questions for the Committee**:*

- *How have (or can) the performance results be used to further the goal of high-quality, efficient healthcare?*
- *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 (4b1. Improvement; 4b2. Benefits of measure)

4b. Usability evaluates 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 has steadily declined from 1.2 percent to 1 percent (January 2015-December 2017 and January 2016-December 2018); however, a slight increase was observed between January 2017 to December 2019 (0.10 percent) and again between January 2018 to December 2020 (0.11 percent).
- The developer also notes a similar decrease in major morbidity during the two analytic periods prior to the start of the COVID-19 pandemic (-0.4 percent) and an increase of 0.69 percent during the same time as mortality.
- The developer reports that the rate of major morbidity has increased from 8.16 percent to 8.85 percent from 2017 to 2019, and 2018 to 2020.

- The developer notes that there may be a few potential explanations the performance including more complete coding of complications by data abstractors as the result of continuing education efforts from STS, the inclusion of unexpected return to the operating room for any reason, as well as the direct and indirect effects of the Covid-19 disease on patients, including disruptions in healthcare that led to many delayed surgeries.

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

Potential harms

- No potential harms noted by the developer.

***Questions for the Committee**:*

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

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

Committee Pre-evaluation Comments:

4a. Use

- Currently publicly reported and used in accountability programs
- The measure is publicly reported. The developer states that it is in use in an accountability program. As of March 2022, approximately 47 percent of STS General Thoracic Surgery Database (GTSD) participants were enrolled in public reporting and receive participant-level results on the following: discharge mortality; median postoperative length of stay for lobectomy procedures for lung cancer; and STS GTSD and National Inpatient Sample (NIS) benchmarks. The developer should identify what feedback it has received from users and how it has responded to the feedback.
- Some feedback is provided, but it is somewhat unclear from the developer submission what feedback was obtained from those being measured and other users and how the feedback is considered when developing or revising the measure specifications
- publicly reported and used in accountability program
- Publicaly reported by 47% of participants enrolled. Graphical illustrations are provided along with a performance report. Developer did not offer feedback input.

4a. Usability

- No concerns
- The developer reports that operative mortality has steadily declined from 1.2 percent to 1 percent (January 2015-December 2017 and January 2016-December 2018); however, a slight increase was observed between January 2017 to December 2019 (0.10 percent) and again between January 2018 to December 2020 (0.11 percent). No potential harms were noted by the developer.
- there are no unintended consequences identified
- no harms

- In current use without unintended consequences.

Criterion 5: Related and Competing Measures

Related/Competing measures

- There are no NQF-endorsed related or competing measures identified by the developer.

Harmonization

- N/A

Committee Pre-evaluation Comments:

5: Related and Competing Measures

- NA
- There are no NQF-endorsed related or competing measures identified by the developer.
- None
- no competing
- n/a

Public and NQF Member Comments

Member Expression of Support

- No members submitted an expression of support for this measure.

Comments

- No NQF member and public comments were received in advance of the Standing Committee evaluation.

Scientific Acceptability Evaluation

RELIABILITY: SPECIFICATIONS

1. **Have measure specifications changed since the last review?** ☐ **Yes** ☒ **No**
2. **Are submitted specifications precise, unambiguous, and complete so that they can be consistently implemented?** ☒ **Yes** ☐ **No**
3. **Briefly summarize any changes to the measure specifications and/or concerns about the measure specifications.**
   - N/A

RELIABILITY: TESTING

4. **Did the developer conduct new reliability testing?** ☐ **Yes** ☒ **No**

   4a. If no, summarize the Standing Committee's previous feedback:

   - The Standing Committee noted previously that the reliability of data elements was supported by an external audit of the GTSD, demonstrating high agreement rates and validation of data accuracy
   - The Standing Committee also noted that the NQF Scientific Methods Panel was satisfied with the reliability testing for the measure.

   4b. If yes, describe any differences between the new and old testing and summarize any relevant Standing Committee's feedback from the previous review:

- N/A

5. **Reliability testing level: ☒ Accountable-Entity Level ☐ Patient/Encounter Level ☐ Neither**
6. **Reliability testing was conducted with the data source and level of analysis indicated for this measure:**

   ☒ Yes ☐ No
7. If accountable-entity level and/or patient/encounter level reliability testing was NOT conducted or if the methods used were NOT appropriate, was **empirical VALIDITY testing** of patient-level data conducted?

   ☐ Yes ☐ No
8. **Assess the method(s) used for reliability testing:**
   - The developer conducted a signal to noise analysis using the STS General Thoracic Surgery Database (GTSD) Version 2.3 (n=233 facilities) and noted ranges from 44.6% (95% credible internal [CrI]= 34.6%-54.1%) to 68% (95% CrI= 53.6%-79.7%).
9. **Assess the results of reliability testing**
   - Providers with at least 30, 50, and 100 cases has reliability scores of 51.7%, 56.1%, and 60.9%, respectively. Large-volume participants (at least 150 cases) has a reliability of 68.0%.
   - The developer notes that while a 0.70 reliability score shows a very close correlation between measured scores, a 0.50 reliability score indicates moderate reliability and demonstrates strong correlation between the true and measures valued of the score.
10. 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.

    ☒ Yes ☐ No ☐ Not applicable
11. Was the method described and appropriate for assessing the reliability of ALL critical data elements?

    ☒ Yes ☐ No ☐ Not applicable (patient/encounter level testing was not performed)
12. **OVERALL RATING OF RELIABILITY** (taking into account precision of specifications and all testing results):

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

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

    ☐ **Low** (NOTE: Should rate LOW if you believe specifications are NOT precise, unambiguous, and complete or if testing methods/results are not adequate)

    ☐ **Insufficient** (NOTE: Should rate INSUFFICIENT if you believe you do not have the information you need to make a rating decision)
13. **Briefly explain rationale for the rating of OVERALL RATING OF RELIABILITY and any concerns you may have with the approach to demonstrating reliability.**
    - Measure specifications precise, unambiguous, and complete (Box 1) -> Empirical reliability testing conducted with the measure as specified (Box 2) -> Reliability testing conducted with computed measure scores (Box 4) -> Method appropriate for assessing variability (signal-to-noise analysis) (Box 5) -> Moderate certainty or confidence that the performance scores are reliable (Box 6a) -> Moderate

VALIDITY: TESTING

14. **Did the developer conduct new validity testing? ☒ Yes ☐ No**

    14a. If no, summarize the Standing Committee's previous feedback:

    14b. If yes, describe any differences between the new and old testing and summarize any relevant Standing Committee's feedback from the previous review:
    - The developer indicates that additional empirical validity testing at the accountable entity level has been conducted since the last review.

- The Standing Committee noted previously that interval testing was performed and only percent agreement was assessed in the analysis. NQF staff clarified that while score-level validity testing is desired, data element testing is acceptable because this is a new measure. For future maintenance evaluations, score-level testing will be required.
- Overall, the Standing Committee did not have any major concerns regarding the validity of the measure and noted that the NQF Scientific Methods Panel was satisfied with the validity analyses for the measure.

15. **Validity testing level (check all that apply):**

    ☒ **Accountable-Entity Level** ☐ **Patient or Encounter-Level** ☐ **Both**

    NOTE: Empirical validity testing is expected at time of maintenance review; if not possible, justification is required.

16. **If patient/encounter level validity testing was provided, was the method described and appropriate for assessing the accuracy of ALL critical data elements? NOTE:** Data element validation from the literature is acceptable.

    ☐ **Yes**

    ☐ **No**

    ☒ **Not applicable** (patient/encounter level testing was not performed)

17. **Method of establishing validity at the accountable-entity level:**

    ☐ **Face validity**

    ☒ **Empirical validity testing at the accountable-entity level**

    ☐ **N/A (accountable-entity level testing not conducted)**

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

    ☒ **Yes**

    ☐ **No**

    ☐ **Not applicable** (accountable-entity level testing was not performed)

19. **Assess the method(s) for establishing validity**

- The developer conducted empirical testing of the composite measure score among three-star rating performance categories (high, average, and low) and assessed composite score stability across two consecutive overlapping reporting periods.
- Providers receiving three stars had lower observed mortality risk (0.4 percent vs. 2.9 percent) and morbidity risk (2.3 percent vs. 20.0 percent) compared to participants receiving one star.
- Composite stability was assessed among 654 participants with at least 10 eligible cases using Pearson's correlation (0.71) and Spearman's correlation (0.74).

20. **Assess the results(s) for establishing validity**

- The developer noted an increase in data accuracy rates and a narrowing of agreement ranges, indicating greater consistency in data accuracy and a high degree of data validity.

VALIDITY: ASSESSMENT OF THREATS TO VALIDITY

21. **Please describe any concerns you have with measure exclusions.**

- N/A

22. **Risk Adjustment**

22a. **Risk-adjustment method**

☐ None (only answer Question 20b and 20e) ☒ Statistical model ☐ Stratification

☐ Other method assessing risk factors (please specify)

22b. **If not risk-adjusted, is this supported by either a conceptual rationale or empirical analyses?**

☐ Yes ☐ No ☒ Not applicable

22c. **Social risk adjustment:**

22c.1 Are social risk factors included in risk model? ☐ Yes ☒ No ☐ Not applicable

22c.2 Conceptual rationale for social risk factors included? ☐ Yes ☐ No

22c.3 Is there a conceptual relationship between potential social risk factor variables and the measure focus? ☐ Yes ☐ No

22d.**Risk adjustment summary:**

22d.1 All of the risk-adjustment variables present at the start of care? ☒ Yes ☐ No

22d.2 If factors not present at the start of care, do you agree with the rationale provided for inclusion? ☐ Yes ☐ No

22d.3 Is the risk adjustment approach appropriately developed and assessed? ☒ Yes ☐ No

22d.4 Do analyses indicate acceptable results (e.g., acceptable discrimination and calibration) ☒ Yes ☐ No

22d.5.Appropriate risk-adjustment strategy included in the measure? ☒ Yes ☐ No

22e**. Assess the risk-adjustment approach**

- The developer used a statistical risk model with 20 risk factors.
- Risk-adjusted operative mortality and major complications rates were estimated using a bivariate random-effects logistic regression model.
- No social risk factors were used in the statistical risk model or for stratification.
- The developer evaluated continuous variables with respect to linearity of effect.
- The model's calibration and discrimination were assessed using Hosmer-Lemeshow statistic and C-statistic.
    - The C-statistic for operative mortality and major morbidity was 0.774 and 0.666, respectively and the developer interpreted these results as the model having good calibration.
    - The Hosmer and Lemeshow Goodness-of-Fit Test results for operative mortality (Chi-square=14.52, degree of freedom [df]= 8, p-value= 0.07 and major morbidity (Chi-Square= 5.54, df=8, p-value=0.07) were interpreted as having good discrimination power and suitable for controlling for differences in case-mix between centers.

23. **Please describe any concerns you have regarding the ability to identify meaningful differences in performance.**

For cost/resource use measures, does this measure identify meaningful differences about cost and resource use between the measured entities?

- The developer determined the degree of uncertainty surrounding a participant's composite measure estimate by calculating a 95% Bayesian credible interval (Cis) and reporting the point estimates and Cis for individual participants along with a comparison to the overall average STS composite score.
- Composite measure results were converted into star categories (n=3); 91.5% of participants (n=140) with at least 30 cases over a three-year period received two stars (performance not statistically significant from overall STS national average).

24. **Please describe any concerns you have regarding comparability of results if multiple data sources or methods are specified.**
    - The measure only uses one set of specifications for this measure.

25. **Please describe any concerns you have regarding missing data.**

- Missing values were imputed when records with missing values of model covariates were identified (except for age, gender, and pathologic stage); patient records missing age, gender, or pathologic state were excluded.
- Values were imputed utilizing the median of observed variables within a category and, for binary risk factors, missing values were considered as absence of the risk factor.
- The range of missing values was between 1% and 3.5%; the developer concluded that there was no bias because of systematic missing data.

For cost/resource use measures ONLY:

If not cost/resource use measure, please skip to question 25.

26. **Are the specifications in alignment with the stated measure intent?**

    ☐ **Yes** ☐ **Somewhat** ☐ **No (If "Somewhat" or "No", please explain)**

27. **Describe any concerns of threats to validity related to attribution, the costing approach, carve outs, or truncation (approach to outliers):**

28. **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 accountable-entity level testing has been conducted)

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

    ☐ **Low** (NOTE: Should rate LOW if you believe that there are threats to validity and/or relevant threats to validity were not assessed OR if testing methods/results are not adequate)

    ☐ **Insufficient** (NOTE: For instrument-based measures and some composite measures, testing at both the accountable-entity level and the patient/encounter level is required; if not conducted, should rate as INSUFFICIENT.)

29. **Briefly explain rationale for rating of OVERALL RATING OF VALIDITY and any concerns you may have with the developers' approach to demonstrating validity.**

    - Potential threats empirically assessed (Box 1) -> Empirical validity testing conducted with the measure as specified (Box 2) -> Validity testing conducted with computed measure scores (Box 5) -> Method appropriate for assessing conceptually and theoretically sound hypothesized relationships (Box 6) -> Moderate certainty or confidence that the performance scores are valid indicator of quality (Box 7a) -> Moderate

FOR COMPOSITE MEASURES ONLY: Empirical analyses to support composite construction

30. **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?**

    ☐ **High**

    ☒ **Moderate**

    ☐ **Low**

    ☐ **Insufficient**

31. **Briefly explain rationale for rating of EMPIRICAL ANALYSES TO SUPPORT COMPOSITE CONSTRUCTION**

    - The developer calculated the operative mortality and major complication rates of 186 hospitals performing at least 30 lobectomies over a three-year period to verify the contribution of each domain to the composite construction.

- The results demonstrate a reduction in mortality and major complication rates from one-star (below average) to three-star (above average) participants.
    - Mortality: 2.1 percent (one star); 1.3 percent (two star); 1.2 percent (three star)
    - Major complications: 16.2 percent (one star); 8.4 percent (two star); 3.2 percent (three star)
- Both domains were divided by their respective standard deviations across STS participants and then added together. An expert panel assessed this weighting and determined that it was consistent with their clinical assessment of each domain's relative importance.
    - Risk-standardized mortality relative weight (0.827)
    - Risk-standardized major morbidity weight (0.173)
- A one percent point change in risk-adjusted mortality rate has the same impact as a 4.8 percentage point change in the site's risk-adjusted morbidity rate.
- The developer notes that while risk-adjusted morbidity explains more of the variation in the overall composite score, it does not dominate, and both domains contribute statistical information.

ADDITIONAL RECOMMENDATIONS

32. **If you have listed any concerns in this form, do you believe these concerns warrant further discussion by the multi-stakeholder Standing Committee? If so, please list those concerns below.**
    - N/A

Criteria 1: 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.

Please separate added or updated information from the most recent measure evaluation within each question response in the Importance to Measure and Report: Evidence section. For example:
2021 Submission:
Updated evidence information here.
2018 Submission:
Evidence from the previous submission here.

Importance to Measure and Report is a threshold criterion that must be met in order to recommend a measure for endorsement. All sub-criteria must be met to pass this criterion. See guidance on evidence .

Please include individual entries for each component measure, unless several components were studied together. If a component measure is submitted as an individual performance measure, complete the evidence section as part of that individual measure submission.

1a. Evidence

1a.01. Provide a logic model.

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.

[Response Begins]
Postoperative complications and operative mortality are important negative outcomes associated with lung cancer resection surgery, including lobectomy, the most frequently performed lung resection procedure. The STS lung cancer resection risk model (Fernandez et al, 2016; Broderick SR 2020) 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.
Broderick SR, Grau-Sepulveda M, Kosinski AS, Kurlansky PA, Shahian DM, Jacobs JP, Becker S, DeCamp MM, Seder CW, Grogan EL, Brown LM, Burfeind W, Magee M, Raymond DP, Puri V, Chang AC, Kozower BD. The Society of Thoracic Surgeons Composite Score Rating for Pulmonary Resection for Lung Cancer. Ann Thorac Surg. 2020 Mar;109(3):848-855. Doi: 10.1016/j.athoracsur.2019.08.114. Epub 2019 Nov 2. PMID: 31689407.

[Response Ends]

1a.02. 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. Otherwise, enter "N/A."

Describe how and from whom input was obtained.

[Response Begins]
N/A
[Response Ends]

1a.03. If this measure is derived from intermediate outcome, process, or structure performance measures, including those that are instrument-based, select the type of source for the systematic review of the body of evidence that supports the performance measure. Otherwise, select "N/A."

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.
[Response Begins]
N/A
[Response Ends]

If the evidence is not based on a systematic review, skip to the end of the section and do not complete the repeatable question group below. If you wish to include more than one systematic review, add additional tables by clicking "Add" after the final question in the group.

Evidence – Systematic Reviews Table (Repeatable)
Group 1 – Evidence – Systematic Reviews Table

1a.04. Provide the title, author, date, citation (including page number) and URL for the systematic review.

[Response Begins]
[Response Ends]

1a.05. Quote the guideline or recommendation verbatim about the process, structure or intermediate outcome being measured. If not a guideline, summarize the conclusions from the systematic review.

[Response Begins]
[Response Ends]

1a.06. Provide the grade assigned to the evidence associated with the recommendation, and include the definition of the grade.

[Response Begins]
[Response Ends]

1a.07. Provide all other grades and definitions from the evidence grading system.

[Response Begins]
[Response Ends]

1a.08. Provide the grade assigned to the recommendation, with definition of the grade.

[Response Begins]
[Response Ends]

1a.09. Provide all other grades and definitions from the recommendation grading system.

[Response Begins]
[Response Ends]

1a.10. Detail the quantity (how many studies) and quality (the type of studies) of the evidence.

[Response Begins]
[Response Ends]

1a.11. Provide the estimates of benefit, and consistency across studies.

[Response Begins]
[Response Ends]

1a.12. Indicate what, if any, harms were identified in the study.

[Response Begins]
[Response Ends]

1a.13. Identify any new studies conducted since the systematic review, and indicate whether the new studies change the conclusions from the systematic review.

[Response Begins]
[Response Ends]

1a.14. Provide empirical data demonstrating the relationship between the outcome (or PRO) and at least one healthcare structure, process, intervention, or service.

[Response Begins]
Data in the STS General Thoracic Surgery Database (GTSD) has demonstrated a reduction in perioperative morbidity and equivalent long-term survival when minimally invasive approaches for lobectomy are used instead of a standard thoracotomy. Specifically, STS data have shown that minimally invasive lung cancer resection has a 50% reduction in major complications compared with a thoracotomy approach, adjusted for age, sex, and comorbidities. There is a general consensus among STS surgeons and the STS GTSD task force that stage I lung cancer is usually resectable with a minimally invasive approach. Because many patients desire a minimally invasive approach, and STS data and other published data demonstrate improved risk-adjusted outcomes, the STS considers it appropriate to include the percent of minimally invasive lobectomies for stage I lung cancer as a process measure on STS biannual reports to GTSD participants. In 2020, Broderick and colleagues reported that "There is wide variability among participants in application of minimally invasive approaches."
As discussed above, the STS lung cancer composite score is based on 2 outcomes: risk-adjusted mortality and morbidity. In 2020, Broderick and colleagues analyzed data from STS GTSD for operations performed from January 2015 to December 2017. "Star ratings" were created for centers with 30 or more cases by using the 95% Bayesian credible intervals. The Bayesian model was performed with and without inclusion of the minimally invasive approach to assess the impact of the approach on the composite measure.
The study population included 38,461 patients from 256 centers. Overall operative mortality was 1.3% (495 of 38,461). The major complication rate was 7.9% (3045 of 38,461). The median number of nodes examined was 10 (interquartile range, 5 to 16); the median number of nodal stations sampled was 4 (interquartile range, 3 to 5). Positive resection margins were identified in 3.7% (1420 out of 38,461). A total of 214 centers with 30 or more cases were assigned star ratings. There were 7 1-star, 194 2-star, and 13 3-star programs; 70.6% of resections were performed through a minimally invasive approach. Inclusion of a minimally invasive approach, which was adjusted for in previous models, altered the star ratings for 3% (6 of 214) of the programs.
References:
Kozower BD, O'Brien SM, Kosinski AS, et al. The Society of Thoracic Surgeons composite score for rating program performance for lobectomy for lung cancer. Ann Thorac Surg 2016;101:1379-87.

Broderick SR, Grau-Sepulveda M, Kosinski AS, Kurlansky PA, Shahian DM, Jacobs JP, Becker S, DeCamp MM, Seder CW, Grogan EL, Brown LM, Burfeind W, Magee M, Raymond DP, Puri V, Chang AC, Kozower BD. The Society of Thoracic Surgeons Composite Score Rating for Pulmonary Resection for Lung Cancer. Ann Thorac Surg. 2020 Mar;109(3):848-855. Doi: 10.1016/j.athoracsur.2019.08.114. Epub 2019 Nov 2. PMID: 31689407.

[Response Ends]

1a.15. If source of evidence is NOT from a clinical practice guideline, USPSTF, or systematic review, describe the evidence on which you are basing the performance measure.

[Response Begins]
Please see 1a.14
[Response Ends]

1a.16. Briefly synthesize the evidence that supports the measure.

[Response Begins]
N/A
[Response Ends]

1a.17. Detail the process used to identify the evidence.

[Response Begins]
N/A
[Response Ends]

1a.18. Provide the citation(s) for the evidence.

[Response Begins]
N/A
[Response Ends]

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

1b.01. Briefly explain the rationale for this measure.

Explain how the measure will improve the quality of care, and list the benefits or improvements in quality envisioned by use of this measure.

[Response Begins]
N/A
[Response Ends]

1b.02. Provide performance scores on the measure as specified (current and over time) at the specified level of analysis.

Include mean, std dev, min, max, interquartile range, and 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 (4b) under Usability and Use.

[Response Begins]
The measure was calculated in two overlapping 3-year time periods; January 1, 2017 – December 31, 2019 and January 1, 2018 – December 31, 2020. For each time period, we provide the number of measured entities (No. of participants), the number of eligible patient records (No. of operations), and the distribution of composite score estimates by percentiles

and geographic region. We present results for all the participants and for the subset of participants with at least 30 eligible cases.

*	January 2017-December 2019	*	January 2018-December 2020	*
*	All Participants	>=30 cases	All Participants	>=30 cases
No. of participants	186	150	176	153
No. of operations	23844	23292	24834	24477
Mean	0.98	0.98	0.979	0.979
SD	0.004	0.005	0.006	0.006
IQR	0.005	0.006	0.008	0.008
Minimum	0.968	0.968	0.952	0.952
10%	0.975	0.974	0.972	0.971
20%	0.977	0.977	0.974	0.974
30%	0.978	0.978	0.977	0.975
40%	0.979	0.979	0.978	0.978
50%	0.98	0.981	0.979	0.979
60%	0.982	0.982	0.981	0.981
70%	0.983	0.983	0.982	0.982
80%	0.984	0.984	0.984	0.984
90%	0.985	0.986	0.985	0.985
Maximum	0.99	0.99	0.989	0.989
Midwest	45	34	41	33
Northeast	48	38	49	43
South	61	53	57	52
West	32	25	29	25

Number of participants, number of operations and the distribution of composite score estimates by percentiles and geographic region

* Indicates that the cell is left intentionally blank

[Response Ends]

1b.03. If no or limited performance data on the measure as specified is reported above, 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. Include citations.

[Response Begins]
Please see data reported in 1b.02
[Response Ends]

1b.04. 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.

Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included. Include mean, std dev, min, max, interquartile range, and scores by decile. 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 (4b) under Usability and Use.

[Response Begins]

*	***January 2017-December 2019 N=23844***	*	***January 2018-December 2020 N=24834***	*
*	***MEAN (SD)***	***MEDIAN (IQR)***	***MEAN (SD)***	***MEDIAN (IQR)***
Age (in years)	**67.5 (9.4)**	**68 (62-74)**	**67.5 (9.4)**	**68 (62-74)**
*	***COUNT***	***PERCENT***	***COUNT***	***PERCENT***
Gender	*	*	*	*
Male	10302	43.21	10736	43.23
Female	13542	56.79	14098	56.77
Insurance < 65	*	*	*	*
Medicare/Medicaid	2307	28.40	2413	28.65
Commercial/HMO	5312	65.39	5496	65.25
None/Self Paid	191	2.35	205	2.43
Other	314	3.87	309	3.67
Insurance >=65	*	*	*	*
Medicare+Medicaid	805	5.12	810	4.94
Medicare+Commercial without Medicaid	8847	56.28	9149	55.75
Medicare without Medicaid/Commercial	6068	38.60	6452	39.32
Race	*	*	*	*
Caucasian	19680	82.54	20507	82.58
Black	1884	7.90	2047	8.24
Hispanic	724	3.04	728	2.93
Asian	960	4.03	952	3.83
Other	316	1.33	328	1.32
Race not available	280	1.17	272	1.10
*	*	*	*	*

Disparities data by different population groups

* Indicates that the cell is left intentionally blank

[Response Ends]

1b.05. If no or limited data on disparities from the measure as specified is reported above, 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 above.

[Response Begins]
Please see data reported in 1b.04
[Response Ends]

1c. Composite- Quality Construct and Rationale

1c.01. Select the method of composite measure construction.

A composite performance measure is a combination of two or more component measures, each of which individually reflect quality of care, into a single performance measure with a single score. For purposes of NQF measure submission, evaluation, and endorsement, the following will be considered composites:

- *Measures with two or more individual performance measure scores combined into one score for an accountable entity.*
- *Measures with two or more individual component measures assessed separately for each patient and then aggregated into one score for an accountable entity.*

- all-or-none measures (e.g., all essential care processes received, or outcomes experienced, by each patient)

[Response Begins]
two or more individual performance measure scores combined into one score
[Response Ends]

1c.02. Describe the quality construct.

Describe the area of quality measured, component measures, and the relationship of the component measures to the overall composite and to each other (whether reflective or formative model was used to develop this measure, and whether components are correlated).

[Response Begins]
The STS Lobectomy Composite Score measures surgical performance for patients treated with lobectomy for lung cancer. Similar to other STS composite measures, this measure is based on a combination of an operative mortality outcome measure and the risk-adjusted occurrence of any of several major complications. To assess overall quality, the composite comprises the following two domains:

1. Operative Mortality (death during the same hospitalization as surgery or within 30 days of the procedure)
2. Presence of at least one of these major complications: pneumonia, acute respiratory distress syndrome, bronchopleural fistula, pulmonary embolus, initial ventilator support greater than 48 hours, reintubation/respiratory failure, tracheostomy, myocardial infarction, or unexpected return to the operating room.

Participants receive a score for each of the two domains, plus an overall composite score. The overall composite score was created by a weighted combination of the above two domains. In addition to receiving a numeric score, participants are assigned to rating categories designated by one to three stars:

1 star: lower-than expected performance
2 stars: as-expected-performance
3 start: higher-than-expected-performance
[Response Ends]

1c.03. Describe the rationale for constructing a composite measure, including how the composite provides a distinctive or additive value over the component measures individually.

[Response Begins]
Risk-adjusted mortality has historically been the dominant outcomes metric for thoracic surgery, but in an era when the average mortality rates for these procedures have declined to very low levels, it can be difficult to differentiate performance based on mortality alone. Specifically, mortality alone fails to take into account the fact that not all operative survivors received equal quality care, e.g., patients who survive surgery but are debilitated by a major postoperative complication. Calculating a composite score from a weighted combination of operative mortality and major complications provides a more comprehensive measure of overall surgical quality.
[Response Ends]

1c.04. Describe how the aggregation and weighting of the component measures are consistent with the stated quality construct and rationale.

[Response Begins]
The composite score is created by a weighted combination of two domains (operative mortality and major complications) resulting in a single composite score. Operative mortality is weighted approximately four times that of a major complication in the composite, consistent with the STS adult cardiac surgery quality measures. The STS General Thoracic Surgery Database working group believes this is an improvement from its previous lung cancer resection model in which mortality and major morbidity were weighted equally.

For more information on the STS composite methodology, please see the attachment:

Kozower BD, O'Brien SM, Kosinski AS, et al. The Society of Thoracic Surgeons composite score for rating program performance for lobectomy for lung cancer. Ann Thorac Surg 2016;101:1379-87.
[Response Ends]

1ma.01. Indicate whether there is new evidence about the measure since the most recent maintenance evaluation. If yes, please briefly summarize the new evidence, and ensure you have updated entries in the Evidence section as needed.

[Response Begins]
Yes
Please see Evidence section for details.
[Response Ends]

Criteria 2: 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.

---

spma.01. Indicate whether there are changes to the specifications since the last updates/submission. If yes, update the specifications in the Measure Specifications section of the Measure Submission Form, and explain your reasoning for the changes below.

[Response Begins]
No
[Response Ends]

spma.02. Briefly describe any important changes to the measure specifications since the last measure update and provide a rationale.

For annual updates, please explain how the change in specifications affects the measure results. If a material change in specification is identified, data from re-testing of the measure with the new specifications is required for early maintenance review.

For example, specifications may have been updated based on suggestions from a previous NQF CDP review.

[Response Begins]
N/A
[Response Ends]

sp.01. Provide the measure title.

Measure titles should be concise yet convey who and what is being measured (see What Good Looks Like).

[Response Begins]
STS Lobectomy for Lung Cancer Composite Score
[Response Ends]

sp.02. Provide a brief description of the measure.

Including type of score, measure focus, target population, timeframe, (e.g., Percentage of adult patients aged 18-75 years receiving one or more HbA1c tests per year).

[Response Begins]
The STS Lobectomy Composite Score comprises two domains:

1. Operative Mortality (death during the same hospitalization as surgery or within 30 days of the procedure)
2. Presence of at least one of these major complications: pneumonia, acute respiratory distress syndrome, bronchopleural fistula, pulmonary embolus, initial ventilator support greater than 48 hours, reintubation/respiratory failure, tracheostomy, myocardial infarction, or unexpected return to the operating room.

The composite score is created by a weighted combination of the above two domains resulting in a single composite score. In addition to receiving a numeric score, participants are assigned to rating categories designated by the following:

1 star: lower-than expected performance
2 stars: as-expected-performance
3 start: higher-than-expected-performance
[Response Ends]

sp.04. Check all the clinical condition/topic areas that apply to your measure, below.

Please refrain from selecting the following answer option(s). We are in the process of phasing out these answer options and request that you instead select one of the other answer options as they apply to your measure.
Please do not select:

- *Surgery: General*

[Response Begins]
Cancer
Surgery: Thoracic Surgery
[Response Ends]

sp.05. Check all the non-condition specific measure domain areas that apply to your measure, below.

[Response Begins]
Other (specify)
N/A
[Response Ends]

sp.06. Select one or more target population categories.

Select only those target populations which can be stratified in the reporting of the measure's result.
Please refrain from selecting the following answer option(s). We are in the process of phasing out these answer options and request that you instead select one of the other answer options as they apply to your measure.
Please do not select:
33. Populations at Risk: Populations at Risk

[Response Begins]
Adults (Age >= 18)
[Response Ends]

sp.07. Select the levels of analysis that apply to your measure.

Check ONLY the levels of analysis for which the measure is SPECIFIED and TESTED.
Please refrain from selecting the following answer option(s). We are in the process of phasing out these answer options and request that you instead select one of the other answer options as they apply to your measure.
Please do not select:

- *Clinician: Clinician*
- *Population: Population*

[Response Begins]
Facility
[Response Ends]

sp.08. Indicate the care settings that apply to your measure.

Check ONLY the settings for which the measure is SPECIFIED and TESTED.
[Response Begins]
Inpatient/Hospital
[Response Ends]

sp.09. 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. If no URL is available, indicate "none available".

[Response Begins]
https://www.sts.org/sites/default/files/STSThoracicDataSpecifications_v5_21_1.pdf
[Response Ends]

sp.11. Attach the data dictionary, code table, or value sets (and risk model codes and coefficients when applicable). Excel formats (.xlsx or .csv) are preferred.

Attach an excel or csv file; if this poses an issue, contact staff . Provide descriptors for any codes. Use one file with multiple worksheets, if needed.
[Response Begins]
No data dictionary/code table – all information provided in the submission form
[Response Ends]

Please respond to the following questions about the numerator, denominator, and exclusions to describe the composite measure, as opposed to the individual component measures.

sp.12. State the numerator.

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.

[Response Begins]
The STS Lobectomy Composite Score comprises two domains:

1. Operative Mortality (death during the same hospitalization as surgery or within 30 days of the procedure)
2. Presence of at least one of these major complications: pneumonia, acute respiratory distress syndrome, bronchopleural fistula, pulmonary embolus, initial ventilator support greater than 48 hours, reintubation/respiratory failure, tracheostomy, myocardial infarction, or unexpected return to the operating room.

The composite score is created by a weighted combination of the above two domains resulting in a single composite score. Operative mortality and major complications were weighted inversely by their respective standard deviations across participants. This procedure is equivalent to first rescaling mortality and complications by their respective standard deviations and then assigning equal weighting to the rescaled mortality rate and rescaled complication rate. This is the same methodology used for other STS composite measures.

In addition to receiving a numeric score, participants are assigned to rating categories designated by the following:

1 star: lower-than expected performance
2 stars: as-expected-performance
3 start: higher-than-expected-performance

Patient Population: The STS GTSD was queried for all patients treated with lobectomy for lung cancer between January 1, 2014, and December 31, 2016. We excluded patients with non-elective status, occult or stage 0 tumors, American Society of Anesthesiologists class VI, and with missing data for age, sex, or discharge mortality status.

Time Window: 01/01/2014 - 12/31/2016

Model variables: Variables in the model: age, sex, year of operation, body mass index, hypertension, steroid therapy,

congestive heart failure, coronary artery disease, peripheral vascular disease, reoperation, preoperative chemotherapy within 6 months, cerebrovascular disease, diabetes mellitus, renal failure, dialysis, past smoker, current smoker, forced expiratory volume in 1 second percent of predicted, Zubrod score (linear plus quadratic), American Society of Anesthesiologists class (linear plus quadratic), and pathologic stage.
[Response Ends]

sp.13. Provide details needed to calculate the numerator.

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 sp.11.

[Response Begins]
Number of patients undergoing elective lobectomy 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
[Response Ends]

sp.14. State the denominator.

Brief, narrative description of the target population being measured.

[Response Begins]
Number of patients greater than or equal to 18 years of age undergoing elective lobectomy for lung cancer
[Response Ends]

sp.15. Provide details needed to calculate the denominator.

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 sp.11.

[Response Begins]
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)
Removal of lung, single lobe (lobectomy) (32480)

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
[Response Ends]

sp.16. Describe the denominator exclusions.

Brief narrative description of exclusions from the target population.

[Response Begins]
Patients were excluded with non-elective status, occult or stage 0 tumors, American Society of Anesthesiologists class VI, and with missing data for age, sex, or discharge mortality status.
[Response Ends]

sp.17. Provide details needed to calculate the denominator exclusions.

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 sp.11.

[Response Begins]
Cases removed from calculations if Emergent, Urgent, or Palliative is checked under "Status of Operation"
OR if T0 is checked under Pathological Staging of the Lung / Lung Tumor: PathStageLungT(1540)
OR if VI is checked under ASA Classification: ASA (1470)
Only general thoracic procedures coded as primary lung or primary esophageal cancer are included in measure calculations, so occult carcinoma is effectively excluded.

[Response Ends]

sp.18. Provide all information required to stratify the measure results, if necessary.

Include 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 in the Data Dictionary field.

[Response Begins]
N/A
[Response Ends]

sp.19. Select the risk adjustment type.

Select type. Provide specifications for risk stratification and/or risk models in the Scientific Acceptability section.
[Response Begins]
Statistical risk model
[Response Ends]

sp.20. Select the most relevant type of score.

Attachment: If available, please provide a sample report.
[Response Begins]
Rate/proportion
[Response Ends]

sp.21. Select the appropriate interpretation of the measure score.

Classifies interpretation of score according to whether better quality or resource use is associated with a higher score, a lower score, a score falling within a defined interval, or a passing score
[Response Begins]
Better quality = Lower score
[Response Ends]

sp.22. Diagram or describe the calculation of the measure score as an ordered sequence of steps.

Identify the target population; exclusions; cases meeting the target process, condition, event, or outcome; time period of data, aggregating data; risk adjustment; etc.

[Response Begins]
Target population is patients treated with lobectomy for lung cancer. Patients were excluded with non-elective status, occult or stage 0 tumors, American Society of Anesthesiologists class VI, and with missing data for age, sex, or discharge mortality status. Outcomes were measured in two domains:

1. Operative Mortality (death during the same hospitalization as surgery or within 30 days of the procedure)
2. Presence of at least one of these major complications: pneumonia, acute respiratory distress syndrome, bronchopleural fistula, pulmonary embolus, initial ventilator support greater than 48 hours, reintubation/respiratory failure, tracheostomy, myocardial infarction, or unexpected return to the operating room.

Time window for analysis was between 01/01/2014 and 12/31/2016.

Analysis considered 24,912 patient records across 233 participant sites.

To form the composite, we rescaled the major complication and operative mortality domains by dividing by their respective standard deviations across STS participants and then added the two domains together. This weighting was then assessed by an expert panel to determine if it provided an appropriate reflection of the relative importance of the two domains.

After rescaling, the relative weights in the final composite of risk-standardized mortality and risk-standardized major morbidity were 0.827 and 0.173, respectively. An implication of this weighting is that a 1 percentage point change in a participant´s risk-adjusted mortality rate has the same impact as a 4.8 percentage point change in the site´s risk-adjusted

morbidity rate. Our expert panel concurred that this weighting was consistent with their clinical assessment of each domain's relative importance.
[Response Ends]

sp.23. Indicate the responder for your instrument.

[Response Begins]
Other (specify)
This is not an instrument-based measure.
[Response Ends]

sp.25. If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.

[Response Begins]
N/A
[Response Ends]

sp.26. Identify whether and how proxy responses are allowed.

[Response Begins]
N/A
[Response Ends]

sp.28. Provide the data collection instrument.

[Response Begins]
Available at measure-specific web page URL identified in sp.09
[Response Ends]

sp.29. Select only the data sources for which the measure is specified.

[Response Begins]
Registry Data
[Response Ends]

sp.30. Describe the component measures and composite construction.

Use this section as needed for aggregation and weighting rules, or calculation of individual performance measures if not individually endorsed.

[Response Begins]
Two or more individual performance measure scores combined into one score
The STS Lobectomy Composite Score measures surgical performance for patients treated with lobectomy for lung cancer. Similar to other STS composite measures, this measure is based on a combination of an operative mortality outcome measure and the risk adjusted
occurrence of any of several major complications. To assess overall quality, the composite comprises the following two domains:
1. Operative Mortality (death during the same hospitalization as surgery or within 30 days of the procedure)
2. Presence of at least one of these major complications: pneumonia, acute respiratory distress syndrome, bronchopleural fistula, pulmonary embolus, initial ventilator support greater than 48 hours, reintubation/respiratory failure, tracheostomy, myocardial
infarction, or unexpected return to the operating room.

Participants receive a score for each of the two domains, plus an overall composite score. The overall composite score was created by a weighted combination of the above two domains. In addition to receiving a numeric score, participants are assigned to rating categories designated by one to three stars:
1 star: lower-than-expected performance
2 stars: as-expected-performance
3 start: higher-than-expected-performance

[Response Ends]

2ma.01. Indicate whether additional empirical reliability testing at the accountable entity level has been conducted. If yes, please provide results in the following section, Scientific Acceptability: Reliability - Testing. Include information on all testing conducted (prior testing as well as any new testing).

Please separate added or updated information from the most recent measure evaluation within each question response in the Scientific Acceptability sections. For example:
Current Submission:
Updated testing information here.
Previous Submission:
Testing from the previous submission here.

[Response Begins]
No
[Response Ends]

2ma.02. Indicate whether additional empirical reliability testing at the accountable entity level has been conducted. If yes, please provide results in the following section, Scientific Acceptability: Validity - Testing. Include information on all testing conducted (prior testing as well as any new testing).

Please separate added or updated information from the most recent measure evaluation within each question response in the Scientific Acceptability sections. For example:
Current Submission:
Updated testing information here.
Previous Submission:
Testing from the previous submission here.

[Response Begins]
Yes
[Response Ends]

2ma.03. For maintenance measures in which risk adjustment/stratification has been performed, indicate whether additional risk adjustment testing has been conducted since the most recent maintenance evaluation. This may include updates to the risk adjustment analysis with additional clinical, demographic, and social risk factors.

Please update the Scientific Acceptability: Validity - Other Threats to Validity section.

Note: This section must be updated even if social risk factors are not included in the risk adjustment strategy.

[Response Begins]
No additional risk adjustment analysis included
[Response Ends]

Measure testing must demonstrate adequate reliability and validity in order to be recommended for endorsement.
Testing must be conducted at the composite score level.
If a component measure is submitted as an individual performance measure, the Scientific Acceptability sections must be completed and submitted as part of the individual measure's submission.

• 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.
• All required sections must be completed.
• For composites with outcome and resource use measures, Questions 2b.23-2b.37 (Risk Adjustment) also must be completed.
• If specified for multiple data sources/sets of specifications (e.g., claims and EHRs), Questions 2b.11-2b.13 also must be completed.
• An appendix for supplemental materials may be submitted (see Question 1 in the Additional section), but there is no guarantee it will be reviewed.
• Contact NQF staff with any questions. Check for resources at the
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 the
2021 Measure Evaluation Criteria and Guidance .
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.
2a. Reliability testing 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 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;
AND
If patient preference (e.g., informed decision-making) 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).
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; 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 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), demonstrate that performance results are not biased due to systematic missing data (or differences between responders and non-responders), and how the specified handling of missing data minimizes bias.
2c. For composite performance measures, empirical analyses support the composite construction approach and demonstrate that:
2c1. the component measures fit the quality construct and add value to the overall composite while achieving the related objective of parsimony to the extent possible; and
2c2. the aggregation and weighting rules are consistent with the quality construct and rationale while achieving the related objective of simplicity to the extent possible.
(if not conducted or results not adequate, justification must be submitted and accepted)
Definitions
Reliability testing applies to the computed measure score. Reliability testing of the measure score addresses precision of measurement (e.g., signal-to-noise).
Validity testing applies to the computed measure score. 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.
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.
Patient preference: Patient preference is not a clinical exception to eligibility and can be influenced by provider interventions.
Risk factors that influence outcomes should not be specified as exclusions.
Meaningful differences*:* 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.

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 below.

Please separate added or updated information from the most recent measure evaluation within each question response in the Importance to Scientific Acceptability sections. For example:
2021 Submission:
Updated testing information here.
2018 Submission:
Testing from the previous submission here.

2a. Reliability

2a.01. Select only the data sources for which the measure is tested.

[Response Begins]
Registry Data
[Response Ends]

2a.02. 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).

[Response Begins]
STS General Thoracic Surgery Database Version 2.3
[Response Ends]

2a.03. Provide the dates of the data used in testing.

Use the following format: "MM-DD-YYYY - MM-DD-YYYY"

[Response Begins]
01/01/2014 – 12/31/2016
[Response Ends]

2a.04. Select the levels of analysis for which the measure is tested.

Testing must be provided for all the levels specified and intended for measure implementation, e.g., individual clinician, hospital, health plan.
Please refrain from selecting the following answer option(s). We are in the process of phasing out these answer options and request that you instead select one of the other answer options as they apply to your measure.
Please do not select:

- *Clinician: Clinician*
- *Population: Population*

[Response Begins]
Facility
[Response Ends]

2a.05. List the measured entities 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.

[Response Begins]
The analysis population consisted of all STS records for patients meeting measure inclusion criteria who had their surgery during January 1, 2014 through December 31, 2016. The population included 24,912 patient records from 233 hospitals.
[Response Ends]

2a.06. Identify the number and descriptive characteristics of patients included in the analysis (e.g., age, sex, race, diagnosis), separated by level of analysis and data source; if a sample was used, describe how patients were selected for inclusion in the sample.

If there is a minimum case count used for testing, that minimum must be reflected in the specifications.

[Response Begins]
Includes 24,912 eligible patients. Patient characteristics are below:

Patient characteristics of patients included in the testing and analysis	*
Age (years), mean (SD)	67.3 (9.5)
Male	44.6%
Body Mass Index (kg/m2), mean, (SD)	27.6 (6.1)
Hypertension	62.0%
Steroid therapy	3.0%
Congestive heart failure	2.5%
Coronary artery disease	20.6%
Peripheral vascular disease	8.9%
Reoperation	5.5%
Preoperative chemotherapy within 6 months	6.5%
Cerebrovascular disease	7.6%
Diabetes mellitus	18.7%
Renal failure	1.1%
Dialysis	0.5%
Cigarette smoking	
Never smoked	15.3%
Past smoker	61.7%
Current smoker	23.0%

Patient characteristics of patients included in the testing and analysis	*
Forced expiratory volume in 1 second percent of predicted	84.5 (19.7)
Zubrod score	*
0	45.9%
1	50.2%
2	3.2%
3	0.6%
4	0.1%
5	<0.1%
ASA Class	*
0	0.2%
2	15.2%
3	76.3%
4	8.3%
5	<0.1%
Pathologic stage	*
0	71.0%
I	17.1%
II	10.4%
IV	1.5%
Year of operation	*
2014	32.1%
2015	34.1%
2016	33.8%

* indicates that cell is left intentionally blank

[Response Ends]

2a.07. 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.

[Response Begins]
The STS tests reliability based on three years of data in the General Thoracic Surgery Database (see 2a.05 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.
[Response Ends]

2a.08. List 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.

[Response Begins]
Patient social risk data is 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.
[Response Ends]

Note: Current guidance for composite measure evaluation states that reliability must be demonstrated for the composite performance measure score.

Note: If accuracy/correctness (validity) of data elements was empirically tested, separate reliability testing of data elements is not required – in 2a.07 check patient or encounter-level data; in 2a.08 enter "see validity testing section of data elements"; and enter "N/A" for 2a.09 and 2a.10.

2a.09. Select the level of reliability testing conducted.

Choose one or both levels.
[Response Begins]
Accountable Entity Level (e.g., signal-to-noise analysis)
[Response Ends]

2a.10. For each level of reliability testing 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.

[Response Begins]
Reliability is conventionally defined as the proportion of variation in a measure that is due to true between-unit differences (i.e., signal) as opposed to random statistical fluctuations (i.e., noise). Equivalently, it is the squared correlation between a measurement and the true value. 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, $\vartheta 1, \ldots, \vartheta N$, 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

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

with $\theta_j^{(l)}$ denoting the value of θ_j 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 $\rho_{(l)}^2$ across the 5,000 MCMC samples.

[Response Ends]

2a.11. For each level of reliability testing checked above, what were the statistical results from reliability testing?

For example, provide the percent agreement and kappa for the critical data elements, or distribution of reliability statistics from a signal-to-noise analysis. For score-level reliability testing, when using a signal-to-noise analysis, more than just one overall statistic should be reported (i.e., to demonstrate variation in reliability across providers). If a particular method

yields only one statistic, this should be explained. In addition, reporting of results stratified by sample size is preferred (pg. 18, NQF Measure Evaluation Criteria).

[Response Begins]
Based on all the 233 participants the reliability (proportion of signal variation) is 44.6%, 95% credible interval [CrI] (34.6%, 54.1%). Reliability increases when considering participants with a particular minimum number of cases within the time window as displayed below.

*	**No Minimum**	**≥30 cases**	**≥50 cases**	**≥100 cases**	**≥150 cases**
No. of participants	233	186	156	101	53
Reliability	44.6%	51.7%	56.1%	60.9%	68.0%
95% CrI	(34.6%-54.1%)	(41.3%-61.4%)	(45.2%-65.6%)	(49.0%-71.2%)	(53.6%-79.7%)

Table showing how reliability increases when considering participants with a particular minimum number of cases within the time window

* indicates that the cell is left intentionally blank

[Response Ends]

2a.12. Interpret the results, in terms of how they demonstrate reliability.

(In other words, what do the results mean and what are the norms for the test conducted?)

[Response Begins]
Reliability increases when considering participants with increasing minimum number of cases. Starting with participants with at least 30 cases, there is a moderate reliability of 0.517 (51.7%), and reliability is 0.68 (68%) when only large-volume participants (at least 150 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

$$\text{Measured Score}_i = \text{True Score}_i + e_i$$

where

$$i = 1,2,\ldots,233$$

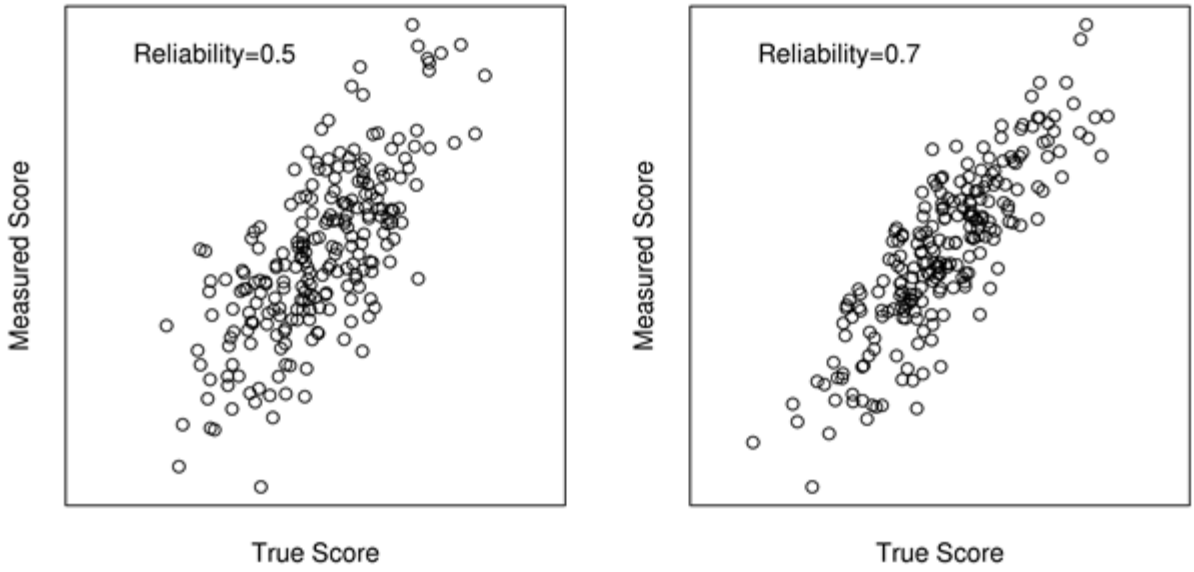
indicates the 233 participants and where

$$\text{True Score}_i$$

and participant error

$$e_i$$

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.



[Response Ends]

2b. Validity

2b.01. Select the level of validity testing that was conducted.

[Response Begins]
Empirical Validity Testing of the Composite (Measure Score)
[Response Ends]

2b.02. 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.

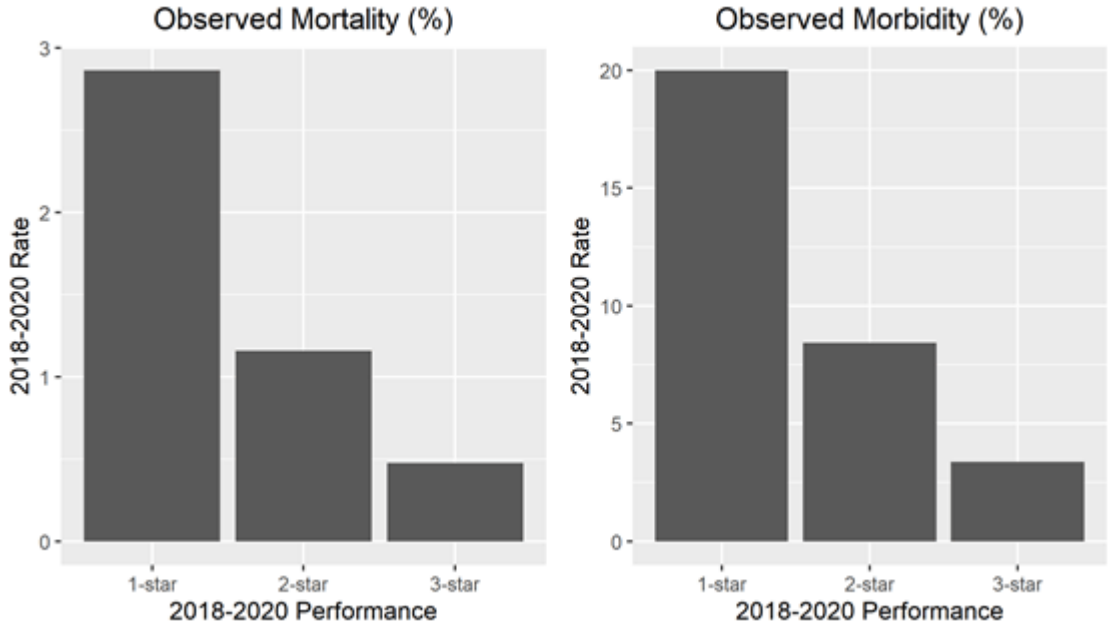
[Response Begins]
The tests on validity used the concept of performance categories to be more formally introduced in 2b4: Participants were labeled as having higher-than-expected performance if the 95% credible interval surrounding a participant's composite score fell entirely above the overall STS average composite score. Participants were labeled as having lower-than-expected performance if the 95% credible interval surrounding a participant's composite score fell entirely below the overall STS average composite score. Participants were labeled as higher-than-expected performance (3 stars), lower-than-expected performance (1 star), and indistinguishable from the average or as-expected performance (2 stars).
We assessed the extent to which a participant's composite score remains stable across two consecutive overlapping reporting periods. This analysis was restricted to 155 participants who participated in each of two consecutive reporting periods: January 2017-December 2020 – January 2018 and December 2021.

[Response Ends]

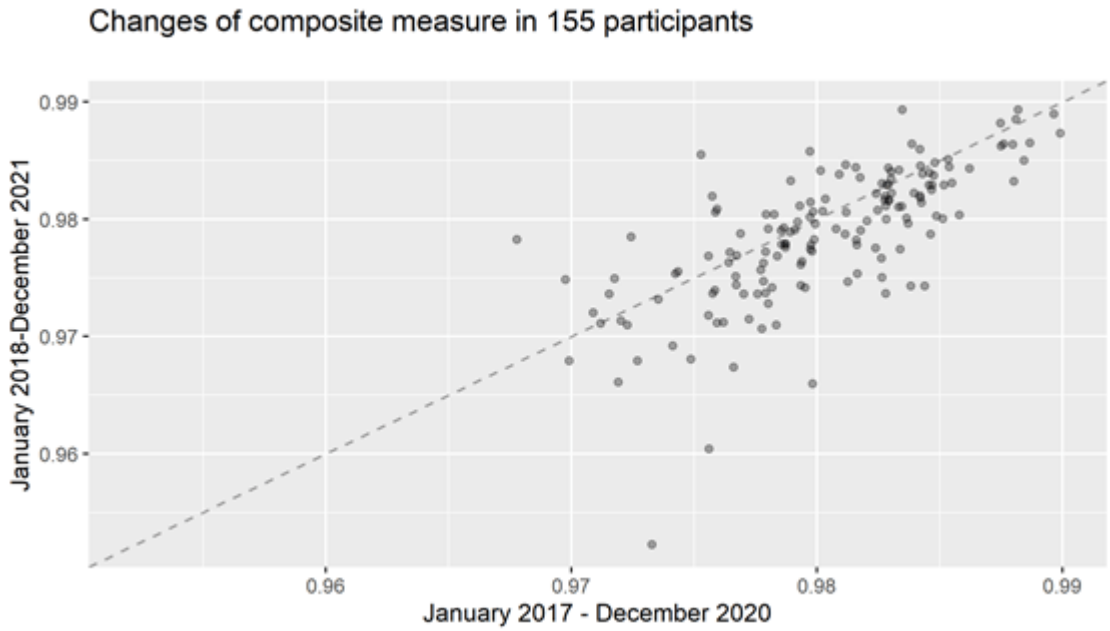
2b.03. Provide the statistical results from validity testing.

Examples may include correlations or t-test results.

[Response Begins]
Compared to participants receiving 1 star, those with 3 stars had lower observed mortality risk (2.9% vs. 0.4%) and lower observed morbidity risk (20.0% vs. 2.3%) during January 2018 – December 2020. Thus, differences in performance were clinically meaningful as well as statistically significant. STS participants deemed better by the composite scores have (on average) higher performance during the same time window on each individual domain of the composite measure.



Stability of the composite measure over time was assessed in 654 participants who participated and had at least 10 eligible cases in each of two consecutive reporting periods: January 2017-December 2020 and January 2018-December 2021.



The Pearson's correlation of the measure between the two time periods is 0.71, the Spearman's correlation is 0.74.

[Response Ends]

2b.04. Provide 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?)

[Response Begins]
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.
[Response Ends]

Note: Applies to the composite performance measure.

2b.05. 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 in Importance to Measure and Report: Gap in Care/Disparities.

[Response Begins]
The degree of uncertainty surrounding an STS participant's composite measure estimate is indicated by calculating 95% Bayesian credible intervals (CI's) which are similar to conventional confidence intervals. Point estimates and CI's for an individual STS participant are reported along with a comparison to the overall average STS composite score. In addition, the composite measure result is converted into categories labeled as 1 to 3 stars. An STS participant receives 2 stars if the Bayesian credible interval surrounding their composite score overlaps the overall STS average. This rating implies that the STS participant's performance was not statistically different from the overall STS national average. If the Bayesian CI falls entirely above the STS national average, the participant receives 3 stars (higher-than-expected performance). If the Bayesian CI falls entirely below the STS national average, the participant receives 1 star (lower-than-expected performance).
[Response Ends]

2b.06. Describe the statistical results from testing the ability to identify statistically significant and/or clinically/practically meaningful differences in performance measure scores across measured entities.

Examples may include 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.

[Response Begins]
Among participants with at least 30 cases over 3 years, 91.5% of participants have received 2 stars, and the remaining participants have received either 1 or 3 stars.
January 1, 2018 through December 31, 2020

*	All Participants	Participants N≥ 30
Category	Number of Participants, %	Number of Participants, %
1-star	4, 2.3%	4, 2.6%
2-star	163, 92.6%	140, 91.5%
3-star	9, 5.1%	9, 5.9%

* indicates that cell is intentionally left blank

[Response Ends]

2b.07. Provide 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.

In other words, what do the results mean in terms of statistical and meaningful differences?

[Response Begins]
The Bayesian methodology allows direct probability interpretation of the results. The identified differences in performance are both statistically significant and clinically meaningful. The surgeon panel and users are satisfied with the distribution of participants across performance categories.
[Response Ends]

Note: Applies to the overall composite measure.

2b.08. Provide the overall frequency of missing data, the distribution of missing data across providers, and the results from testing related to missing data.

For example, provide results of sensitivity analysis of the effect of various rules for missing data/non-response. If no empirical sensitivity analysis was conducted, identify the approaches for handling missing data that were considered and benefits and drawbacks of each).

[Response Begins]
To maximize use of available data, when encountering records with missing values of model covariates (with the exception of age and gender and pathologic stage), the missing values were imputed. Patient records missing age or gender or pathologic stage were excluded. Variables FEV1 and steroid use were each missing for approximately 3% of patients. Remaining variables had less than 1% of missing values.
[Response Ends]

2b.09. Describe the method of testing conducted to identify the extent and distribution of missing data (or non-response) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and non-responders). Include how the specified handling of missing data minimizes bias.

Describe the steps—do not just name a method; what statistical analysis was used.

[Response Begins]
The quality of data in the STS General Thoracic Surgery Database has been improving. We managed the missing data with imputation. Missing body mass index (BMI) values (0.27%) were imputed utilizing the median of the observed BMI values. Missing FEV1 (2.65%) was imputed to the median within the smoking status categories. For binary risk factors, missing values were considered as indicating absence of the risk factor.
[Response Ends]

2b.10. Provide 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 non-responders), and how the specified handling of missing data minimizes bias.

In other words, 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 was conducted, justify the selected approach for missing data.

[Response Begins]
The rates of missing data were low and are getting lower. We therefore concluded that systematic missing data did not lead to bias in our measure.
[Response Ends]

Note: Applies to all component measures, unless already endorsed or are being submitted for individual endorsement.

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 eCQMs). 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.

2b.11. Indicate whether there is more than one set of specifications for this measure.

[Response Begins]
No, there is only one set of specifications for this measure
[Response Ends]

2b.12. 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. Indicate what statistical analysis was used.

[Response Begins]
N/A
[Response Ends]

2b.13. Provide the statistical results from testing comparability of performance scores for the same entities when using different data sources/specifications.

Examples may include correlation, and/or rank order.

[Response Begins]
N/A
[Response Ends]

2b.14. Provide your interpretation of the results in terms of the differences in performance measure scores for the same entities across the different data sources/specifications.

In other words, what do the results mean and what are the norms for the test conducted.

[Response Begins]
N/A
[Response Ends]

Applies to the composite performance measure, as well all component measures unless they are already endorsed or are being submitted for individual endorsement.

2b.15. Indicate whether the measure uses exclusions.

[Response Begins]
Yes, the measure uses exclusions.
[Response Ends]

2b.16. Describe the method of testing exclusions and what was tested.

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?

[Response Begins]
We excluded patients with missing data for age, sex, or discharge mortality status. In addition we excluded patients with non-elective status, occult or stage 0 tumors, missing pathologic stage or American Society of Anesthesiologists class VI. 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.

[Response Ends]

2b.17. Provide 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.

[Response Begins]
There were 155 (0.6%) occult or stage 0 tumors, 416 (1.6%) with missing pathologic stage 2 (0.008%) ASA VI, and 254 (1.0 %) non-elective status patients, resulting in the overall exclusion of 3.2% (810 of 25640 patient records – final population 24830 records). Impact of these exclusions on the performance measure is negligible due to the small proportion of cases excluded.

[Response Ends]

2b.18. Provide your interpretation of the results, in terms of demonstrating that exclusions are needed to prevent unfair distortion of performance results.

In other words, 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.

[Response Begins]
For the measure to consistently quantify the surgical quality of lobectomy for lung cancer per its definition (outcome domains of operative mortality and major complications), it is necessary and clinically appropriate to exclude cases with non-elective status, missing pathology or occult or stage 0 tumors, or American Society of Anesthesiologists class VI.

[Response Ends]

Note: Applies to all outcome or resource use component measures, unless already endorsed or are being submitted for individual endorsement.

2b.19. Check all methods of controlling for differences in case mix that was used.

[Response Begins]
Statistical risk model with risk factors (specify number of risk factors)
20
[Response Ends]

2b.20. If using statistical risk models, provide detailed risk model specifications, including the risk model method, risk factors, risk factor data sources, coefficients, equations, codes with descriptors, and definitions.

[Response Begins]
Participant-specific risk-adjusted operative mortality and major complication rates were estimated using a bivariate random-effects logistic regression model. The term bivariate refers to the fact that both operative mortality and major complications were analyzed together in a single model, not estimated one at a time in separate models. 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. Detailed description is provided in published statistical appendix; a copy is appended to the end of this document. Risk factors in the model were: age, sex, year of operation, body mass index, hypertension, steroid therapy, congestive heart failure, coronary artery disease, peripheral vascular disease, reoperation, preoperative chemotherapy within 6 months, cerebrovascular disease, diabetes mellitus, renal failure, dialysis, past smoker, current smoker, forced expiratory volume in 1 second percent of predicted,

Zubrod score (linear plus quadratic), American Society of Anesthesiologists class (linear plus quadratic), and pathologic stage.
[Response Ends]

2b.21. If an outcome or resource use measure is not risk-adjusted or stratified, provide rationale and analyses to demonstrate that controlling for differences in patient characteristics (i.e., case mix) is not needed to achieve fair comparisons across measured entities.

[Response Begins]
N/A
[Response Ends]

2b.22. Describe the conceptual and statistical methods and criteria used to test and select patient-level risk factors (e.g., clinical factors, social risk factors) used in the statistical risk model or for stratification by risk.

Please be sure to address the following: potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of p<0.10 or other statistical tests; correlation of x or higher. Patient factors should be present at the start of care, if applicable. Also discuss any "ordering" of risk factor inclusion; note whether social risk factors are added after all clinical factors. Discuss any considerations regarding data sources (e.g., availability, specificity).

[Response Begins]
Covariates in this model were selected a priori based on a combination of literature review and expert group consensus, and as described in Kozower, 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.
Kozower BD, O'Brien SM, Kosinski AS, et al. The Society of Thoracic Surgeons composite score for rating program performance for lobectomy for lung cancer. Ann Thorac Surg 2016;101:1379-87.

[Response Ends]

2b.23. Select all applicable resources and methods used to develop the conceptual model of how social risk impacts this outcome.

[Response Begins]
Published literature
Other (specify)
Expert group consensus
[Response Ends]

2b.24. Detail the statistical results of the analyses used to test and select risk factors for inclusion in or exclusion from the risk model/stratification.

[Response Begins]
Estimated odds ratios are summarized in the table below.

*	Operative Mortality	*	Major Morbidity	*
Variable	**OR (95% CI)**	**p-value**	**OR (95% CI)**	**p-value**
Age, yrs, (per 1 yr increase)	1.069 (1.050, 1.088)	<.0001	1.007 (1.001, 1.013)	0.0159
Male	1.323 (1.010, 1.732)	0.0418	1.348 (1.226, 1.483)	<.0001
Body Mass Index (kg/m2), (per 1 unit increase)	0.972 (0.948, 0.997)	0.026	0.970 (0.962, 0.979)	<.0001
Hypertension	0.938 (0.685, 1.283)	0.6876	1.114 (0.999, 1.241)	0.0515
Steroid therapy	2.037 (1.300, 3.191)	0.0019	1.154 (0.934, 1.426)	0.1845
Congestive heart failure	1.555 (0.985, 2.456)	0.0583	1.374 (1.123, 1.683)	0.0021
Coronary artery disease	0.937 (0.688, 1.275)	0.6781	0.984 (0.874, 1.108)	0.7928
Peripheral vascular disease	1.645 (1.196, 2.261)	0.0022	1.154 (1.005, 1.324)	0.0418
Reoperation	2.349 (1.593, 3.463)	<.0001	1.319 (1.089, 1.597)	0.0046
Preoperative chemotherapy within 6 months	1.584 (0.986, 2.543)	0.0571	1.317 (1.095, 1.583)	0.0035
Cerebrovascular disease	1.079 (0.749, 1.556)	0.6822	1.148 (0.994, 1.327)	0.0612
Diabetes mellitus	1.296 (0.963, 1.745)	0.0869	1.008 (0.896, 1.133)	0.8987
Renal failure	0.913 (0.330, 2.529)	0.8614	0.973 (0.628, 1.507)	0.9027
Dialysis	1.668 (0.399, 6.977)	0.4832	2.013 (1.166, 3.474)	0.012
Past smoker	1.645 (0.984, 2.749)	0.0578	1.479 (1.256, 1.742)	<.0001
Current smoker	2.197 (1.257, 3.841)	0.0058	2.001 (1.677, 2.388)	<.0001
FEV in 1 second percent of predicted (per 1 unit increase)	0.991 (0.984, 0.998)	0.0139	0.990 (0.987, 0.992)	<.0001
Zubrod score (per 1 unit increase)	1.848 (1.278, 2.672)	0.0011	1.237 (1.073, 1.426)	0.0034
Squared Zubrod score (per 1 unit increase)	0.942 (0.812, 1.094)	0.436	1.036 (0.971, 1.106)	0.2876
ASA Class (per 1 unit increase)	19.439 (1.126, 335.665)	0.0412	1.742 (0.858, 3.539)	0.1245
Squared ASA Class (per 1 unit increase)	0.683 (0.444, 1.049)	0.0813	0.956 (0.852, 1.073)	0.448
Pathologic stage I	1.157 (0.845, 1.582)	0.3628	1.147 (1.025, 1.283)	0.0166
Pathologic stage II	1.555 (1.098, 2.202)	0.0129	1.271 (1.112, 1.454)	0.0004
Pathologic stage IV	2.223 (0.988, 5.001)	0.0535	1.545 (1.096, 2.176)	0.0129
Year of operation (per 1 yr increase)	1.008 (0.858, 1.183)	0.9267	1.160 (1.095, 1.229)	<.0001

* indicates that the cell is intentionally left blank

[Response Ends]

2b.25. Describe the analyses and interpretation resulting in the decision to select or not select social risk factors.

Examples may include prevalence of the factor across measured entities, availability of the data source, empirical association with the outcome, contribution of unique variation in the outcome, or assessment of between-unit effects and within-unit effects. Also describe the impact of adjusting for risk (or making no adjustment) on providers at high or low extremes of risk.

[Response Begins]
All covariates were retained in the model and were not added or removed based on a statistical variable selection algorithm.
Patient social risk data are not collected in the General Thoracic Surgery Database.

[Response Ends]

2b.26. 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). Provide the statistical results from testing the approach to control for differences in patient characteristics (i.e., case mix) below. If stratified ONLY, enter "N/A" for questions about the statistical risk model discrimination and calibration statistics.

Validation testing should be conducted in a data set that is separate from the one used to develop the model.

[Response Begins]
Continuous variables were evaluated with respect to linearity of effect and needed transformations were considered resulting in addition of squared ASA class and Zubrod score. The calibration of the model was assessed with the Hosmer-Lemeshow statistic. The discrimination of the model was assessed with the C-statistic.
[Response Ends]

2b.27. Provide risk model discrimination statistics.

For example, provide c-statistics or R-squared values.

[Response Begins]
Operative mortality model: C-statistic is 0.774. Major morbidity model: C-statistic is 0.666
[Response Ends]

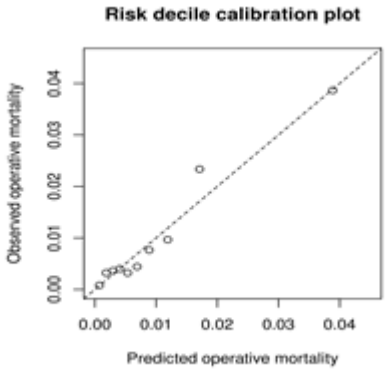
2b.28. Provide the statistical risk model calibration statistics (e.g., Hosmer-Lemeshow statistic).

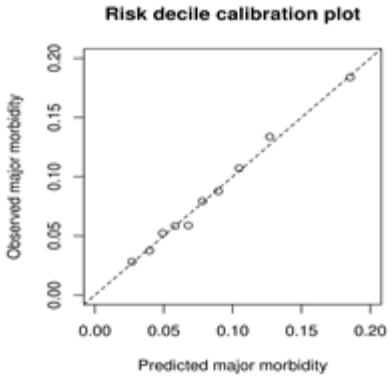
[Response Begins]
Operative mortality model: Hosmer and Lemeshow Goodness-of-Fit Test p-value=0.07 (Chi-Square=14.52, df=8). Major morbidity model: Hosmer and Lemeshow Goodness-of-Fit Test p-value=0.70 (Chi-Square=5.54, df=8).
[Response Ends]

2b.29. Provide the risk decile plots or calibration curves used in calibrating the statistical risk model.

The preferred file format is .png, but most image formats are acceptable.

[Response Begins]
Risk decile plots below show good alignment of predicted and observed probabilities of outcome (operative mortality and major morbidity) within deciles of predicted values.





[Response Ends]

2b.30. Provide the results of the risk stratification analysis.

[Response Begins]
N/A
[Response Ends]

2b.31. Provide your interpretation of the results, in terms of demonstrating adequacy of controlling for differences in patient characteristics (i.e., case mix).

In other words, what do the results mean and what are the norms for the test conducted?

[Response Begins]
The results demonstrated that the STS lobectomy risk models are well calibrated and have good discrimination power. They are suitable for controlling for differences in case-mix between centers.
[Response Ends]

2b.32. Describe any additional testing conducted to justify the risk adjustment approach used in specifying the measure.

Not required but would provide additional support of adequacy of the risk model, e.g., testing of risk model in another data set; sensitivity analysis for missing data; other methods that were assessed.

[Response Begins]
N/A
[Response Ends]

Note: If empirical analyses do not provide adequate results—or are not conducted—justification must be provided and accepted in order to meet the must-pass criterion of Scientific Acceptability of Measure Properties. Each of the following questions has instructions on what to provide if no empirical analysis was conducted.

2c. Composite – Empirical Analysis

2c.01. Provide empirical analysis demonstrating that the component measures fit the quality construct, add value to the overall composite, and achieve the object of parsimony to the extent possible.

[Response Begins]
Please see below.
[Response Ends]

2c.02. Describe the method used to support the composite construction.

Describe the steps—do not just name a method; indicate what statistical analysis was used; if no empirical analysis, provide a justification.

[Response Begins]
To verify that each domain contributes statistical information, we calculated the operative mortality and major complication rates across program star ratings among 186 hospitals with at least 30 lobectomies within three years.
[Response Ends]

2c.03. Provide the statistical results obtained from the analysis of the components.

Examples include correlations, contribution of each component to the composite score, etc.; if no empirical analysis, identify the components that were considered and the pros and cons of each.

[Response Begins]
The table below demonstrates that the mortality and major complication rates decrease monotonically from one-star (below average) to three-star (above average) participants.
Operative Mortality and Major Complication Rates Across Star Ratings

*	One star	Two Star	Three Star	All Programs
Operative mortality (95% CI)	2.1% (1.4%, 3.2%)	1.3% (1.1%, 1.4%)	0.4% (0.2%, 0.7%)	1.2% (1.1%, 1.4%)
Major complication (95% CI)	16.2% (14.1%, 18.6%)	8.4% (8.0%, 8.8%)	3.2% (2.5%, 4.1%)	8.3% (8.0%, 8.7%)

Operative mortality and major complication rates across star ratings
Among 186 hospitals with at least 30 lobectomies.
* indicates that the cell is intentionally left blank

[Response Ends]

2c.04. Provide your interpretation of the results, in terms of demonstrating that the components included in the composite are consistent with the described quality construct and add value to the overall composite.

In other words, what do the results mean in terms of supporting inclusion of the components; if no empirical analysis, provide rationale for the components that were selected.

[Response Begins]
Although risk-adjusted morbidity explains more of the variation in the overall composite score, it does not dominate. Both domains contribute statistical information.
[Response Ends]

2c.05. Provide an empirical analysis demonstrating that the aggregations and weighting rules are consistent with the quality construct and achieve the objective of simplicity to the extent possible.

[Response Begins]
Please see below.
[Response Ends]

2c.06. Describe the method used for composite aggregation.

Describe the steps—do not just name a method; what statistical analysis was used; if no empirical analysis, provide justification.

[Response Begins]
To form the composite, we rescaled the morbidity and mortality domains by dividing by their respective standard deviations across STS participants and then added the two domains together. This weighting was then assessed by an expert panel to determine if it provided an appropriate reflection of the relative importance of the two domains.
[Response Ends]

2c.07. Provide the statistical results obtained from the analysis of the aggregation and weighting rules.

If no empirical analysis was conducted, identify the aggregation and weighting rules that were considered and the pros and cons of each.

[Response Begins]
After rescaling, the relative weights in the final composite of risk-standardized mortality and risk-standardized major morbidity were 0.827 and 0.173, respectively. An implication of this weighting is that a 1 percentage point change in a participant's risk-adjusted mortality rate has the same impact as a 4.8 percentage point change in the site's risk-adjusted morbidity rate.
[Response Ends]

2c.08. Provide your interpretation of the results, in terms of demonstrating the aggregation and weighting rules are consistent with the described quality construct.

In other words, what do the results mean in terms of supporting the selected rules for aggregation and weighting; if no empirical analysis, provide rationale for the selected rules for aggregation and weighting.

[Response Begins]
This weighting was consistent with our expert panel's clinical assessment of each domain's relative importance.
[Response Ends]

Criteria 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.

3.01. Check all methods below that are used to generate the data elements needed to compute the measure score.

[Response Begins]
Generated or collected by and used by healthcare personnel during the provision of care (e.g., blood pressure, lab value, diagnosis, depression score)
Coded by someone other than person obtaining original information (e.g., DRG, ICD-10 codes on claims)
Abstracted from a record by someone other than person obtaining original information (e.g., chart abstraction for quality measure or registry)
[Response Ends]

3.02. Detail to what extent the specified data elements are available electronically in defined fields.

In other words, indicate whether data elements that are needed to compute the performance measure score are in defined, computer-readable fields.
[Response Begins]
ALL data elements are in defined fields in a combination of electronic sources
[Response Ends]

3.03. 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 data elements not from electronic sources.

[Response Begins]
All data elements from STS General Thoracic Surgery Database (GTSD) participating institutions are submitted in electronic format following a standard set of data specifications.
[Response Ends]

3.04. Describe any efforts to develop an eCQM.

[Response Begins]
N/A
[Response Ends]

3.06. 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.

[Response Begins]
There are no known major difficulties. The data elements included in this measure have been standard in the STS GTSD for at least 3 years and some of them have been part of the database for 20 years. The variables are considered to be data elements that are readily available and collected as part of the process of providing care.
The STS GTSD has two lock dates (i.e., data harvest) a year, in which a snapshot of the data are sent from the data warehouse to the analytic center for analyses. Between lock dates, database participants are able to access reports updated in near real time through a number of dashboard reports and tools available in the database platform. Using these reports, database participants are able to address data quality and completeness. Data harvest results are provided to participants via the database platform. Due to COVID-19 and the 2020 STS National Database data warehouse transition, STS experienced some harvest delays, which have been addressed.
Data Collection:

STS GTSD participants have on-staff data managers or use third party abstraction companies to collect these data. Costs to develop the measure included volunteer thoracic surgeons' time, STS staff time, and analytic center statistician and project management time.
Other fees:
STS GTSD participants pay an annual participant fee of $550 or $975 per surgeon depending on whether the participant is an STS member or not.

[Response Ends]

Consider implications for both individuals providing data (patients, service recipients, respondents) and those whose performance is being measured.

3.07. Detail 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),

Attach the fee schedule here, if applicable.

[Response Begins]
Please see 3.06
[Response Ends]

Criteria 4: Use and Usability

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.

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

4a. Use

4a.01.

Check all current uses. For each current use checked, please provide:

Name of program and sponsor

URL

Purpose

Geographic area and number and percentage of accountable entities and patients included

Level of measurement and setting

[Response Begins]
Public Reporting
Quality Improvement with Benchmarking (external benchmarking to multiple organizations)
Quality Improvement (Internal to the specific organization)
As a national leader in health care transparency and accountability, STS believes that the public has a right to know the quality of surgical outcomes. Launched in 2010, STS public reporting is a voluntary initiative in which STS publishes measure results of consenting STS National Database participants on its website – https://publicreporting.sts.org/ . The STS public reporting website is updated with new data once a year.
STS continues to experience a steady increase of its database participants enrolling in public reporting. As of March 2022, approximately 47% of STS General Thoracic Surgery Database (GTSD) participants were enrolled in public reporting. The data reported include participant-level results for discharge mortality and median postoperative length of stay for lobectomy procedures for lung cancer along with STS GTSD and National Inpatient Sample (NIS) benchmarks. In addition, overall and domain-specific scores and star ratings for the Lobectomy for Lung Cancer Composite Measure are reported. GTSD public reporting online may be found here:
http://publicreporting.sts.org/gtsd .
GTSD participants are encouraged to use their feedback report results for internal assessment and quality improvement.

STS GTSD participant feedback reports that include the lobectomy composite scores and star ratings are provided twice a year to GTSD participants who actively submit data. STS and NIS (where appropriate) benchmarks are provided.

[Response Ends]

4a.02. Check all planned uses.

[Response Begins]
Public reporting
Quality Improvement with Benchmarking (external benchmarking to multiple organizations)

Quality Improvement (internal to the specific organization)
[Response Ends]

4a.03. If not currently publicly reported OR used in at least one other accountability application (e.g., payment program, certification, licensing), explain why the measure is not in use.

For example, do policies or actions of the developer/steward or accountable entities restrict access to performance results or block implementation?

[Response Begins]
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
[Response Ends]

4a.04. If not currently publicly reported OR used in at least one other accountability application, provide a credible plan for implementation within the expected timeframes: used in any accountability application within 3 years, and publicly reported within 6 years of initial endorsement.

A 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.

[Response Begins]
n/a
[Response Ends]

4a.05. Describe how performance results, data, and assistance with interpretation have been provided to those being measured or other users during development or implementation.

Detail 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.

[Response Begins]
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 where they perform compared to their peers on a bi-annual basis. In addition, these risk-adjusted results allow surgeons to benchmark their program and initiate QI efforts, as needed.
In providing transparency through public reporting of this measure, surgeons can better compare their patients´ outcomes with national benchmarks and patients will be better informed consumers of health care.

[Response Ends]

4a.06. Describe the process for providing measure results, including when/how often results were provided, what data were provided, what educational/explanatory efforts were made, etc.

[Response Begins]
The performance reports include separate mortality and morbidity domain scores and an overall composite score. The surgeon´s score is illustrated graphically in relation to the 25th, 50th and 75th percentiles of the distribution across all surgeons who were eligible for inclusion in the analysis for the specified three-year period, and is also accompanied by the 95% Bayesian credible interval. A detailed report overview, providing explanations of statistical calculations, endpoints, and report interpretation, is included in the report.
[Response Ends]

4a.07. Summarize the feedback on measure performance and implementation from the measured entities and others. Describe how feedback was obtained.

[Response Begins]
The general thoracic surgeons 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.
[Response Ends]

4a.08. Summarize the feedback obtained from those being measured.

[Response Begins]
See 4a.07
[Response Ends]

4a.09. Summarize the feedback obtained from other users.

[Response Begins]
N/A, no measure-specific feedback received.
[Response Ends]

4a.10. Describe how the feedback described has been considered when developing or revising the measure specifications or implementation, including whether the measure was modified and why or why not.

[Response Begins]
N/A
[Response Ends]

4b. Usability

4b.01. You may refer to data provided in Importance to Measure and Report: Gap in Care/Disparities, 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, provide an explanation. 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.

[Response Begins]
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 (ABTS). 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.
Operative mortality in the STS General Thoracic Surgery Database had a steady decline in the previous years – from 1.2% between January 2015 and December 2017 to 1% between January 2016 and December 2018. Between January 2017 and December 2019, operative mortality rate was 1.10% which has shown a slight increase of 0.10 % and was approximately the same between January 2018 – December 2020 at 1.11%.
Major morbidity rate has increased from 8.16% to 8.85% during the same time. However, the major morbidity rate had a noticeable and a steady decline prior to the start of the Covid-19 pandemic when compared to previous years – 8.4% (January 2015 – December 2017) and 8% (January 2016 – December 2018). A few potential explanations for this observation are more complete coding of complications by data abstractors as the result of continuing education efforts from STS, the inclusion of unexpected return to the operating room for any reason, as well as the direct and indirect effects of the Covid-19 disease on patients, including disruptions in healthcare that led to many delayed surgeries.

[Response Ends]

4b.02. Explain any unexpected findings (positive or negative) during implementation of this measure, including unintended impacts on patients.

[Response Begins]
We are not aware of any unexpected findings associated with implementation of this measure.
[Response Ends]

4b.03. Explain any unexpected benefits realized from implementation of this measure.

[Response Begins]
N/A
[Response Ends]

Criteria 5: Related and Competing Measures

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

If you are updating a maintenance measure submission for the first time in MIMS, please note that the previous related and competing data appearing in question 5.03 may need to be entered in to 5.01 and 5.02, if the measures are NQF endorsed. Please review and update questions 5.01, 5.02, and 5.03 accordingly.

5.01. Search and select all NQF-endorsed related measures (conceptually, either same measure focus or target population).

(Can search and select measures.)
[Response Begins]
[Response Ends]

5.02. Search and select all NQF-endorsed competing measures (conceptually, the measures have both the same measure focus or target population).

(Can search and select measures.)
[Response Begins]
[Response Ends]

5.03. If there are related or competing measures to this measure, but they are not NQF-endorsed, please indicate the measure title and steward.

[Response Begins]
Measure #1790 - Risk-Adjusted Morbidity and Mortality for Lung Resection for Lung Cancer. The measure steward is the Society of Thoracic Surgeons.
[Response Ends]

5.04. If this measure conceptually addresses EITHER the same measure focus OR the same target population as NQF-endorsed measure(s), indicate whether the measure specifications are harmonized to the extent possible.

[Response Begins]
Yes
[Response Ends]

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

[Response Begins]
N/A
[Response Ends]

5.06. Describe why this measure is superior to competing measures (e.g., a more valid or efficient way to measure quality). Alternatively, justify endorsing an additional measure.

Provide analyses when possible.

[Response Begins]
N/A
[Response Ends]

Appendix

Supplemental materials may be provided in an appendix. : No appendix

Contact Information

Measure Steward (Intellectual Property Owner): The Society of Thoracic Surgeons
Measure Steward Point of Contact: Yagci, Banu, byagci@sts.org
Measure Developer if different from Measure Steward: The Society of Thoracic Surgeons
Measure Developer Point(s) of Contact: Yagci, Banu, byagci@sts.org

Additional Information

1. Provide any supplemental materials, if needed, as an appendix. All supplemental materials (such as data collection instrument or methodology reports) should be collated one file with a table of contents or bookmarks. If material pertains to a specific criterion, that should be indicated.

[Response Begins]
No appendix
[Response Ends]

2. List the workgroup/panel members' names and organizations.

Describe the members' role in measure development.

[Response Begins]
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.
Task Force on Quality Initiatives
Gaetano Paone, MD, Chair; Henry Ford Hospital, Detroit, MI
William Burfeind, MD; St. Luke's University Health Network, Bethlehem, PA
William Caine, MD; Intermountain Heart Institute, Murray, UT
Fred Edwards, MD; University of Florida, Jacksonville, FL
Chris Feindel, MD; Cardiovascular Surgery Associates, Toronto, Ontario
Felix Fernandez, MD; Emory University School of Medicine, Atlanta, GA
Kristopher George, MD; Cardiac Surgical Associates of Fresno, Fresno, CA
Fred Grover, MD; University of Colorado School of Medicine, Aurora, CO
Jeffrey P. Jacobs, MD; University of Florida, Gainesville, FL
Kevin Lobdell, MD; Atrium Health, Charlotte, NC
John Mayer, MD; Boston Childrens Hospital, Boston, MA
Jim McClurken, MD; Doylestown Hospital, Doylestown, PA
Edward Savage, MD; Cleveland Clinic/Martin Health, Stuart, FL
David M. Shahian, MD; Massachusetts General Hospital & Harvard Medical School, Boston, MA
Frank Shannon, MD; Johns Hopkins All Children's Hospital, St. Petersburg, FL
Robert Welsh, MD; Beaumont Health, Royal Oak, MI

[Response Ends]

3. Indicate the year the measure was first released.

[Response Begins]
2016
[Response Ends]

4. Indicate the month and year of the most recent revision.

[Response Begins]
April 2021
[Response Ends]

5. Indicate the frequency of review, or an update schedule, for this measure.

[Response Begins]

Annually
[Response Ends]

6. Indicate the next scheduled update or review of this measure.

[Response Begins]
2022
[Response Ends]

7. Provide a copyright statement, if applicable. Otherwise, indicate "N/A".

[Response Begins]
N/A
[Response Ends]

8. State any disclaimers, if applicable. Otherwise, indicate "N/A".

[Response Begins]
N/A
[Response Ends]

9. Provide any additional information or comments, if applicable. Otherwise, indicate "N/A".

[Response Begins]
N/A
[Response Ends]