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 #: 3357
Measure Title: Facility-Level 7-Day Hospital Visits after General Surgery Procedures Performed at Ambulatory Surgical Centers
Measure Steward: Centers for Medicare & Medicaid Services (CMS)
Brief Description of Measure: Facility-level risk-standardized rate of acute, unplanned hospital visits within 7 days of a general surgery procedure performed at an ambulatory surgical center (ASC) among Medicare Fee-For-Service (FFS) patients aged 65 years and older. An unplanned hospital visit is defined as an emergency department (ED) visit, observation stay, or unplanned inpatient admission.
Developer Rationale: This measure aims to reduce adverse patient outcomes associated with ASC surgeries and improve follow-up care by capturing and illuminating, for providers and patients, post-surgery unplanned hospital visits that are often not visible to providers at ASCs. The measure score will assess quality and inform quality improvement.

Numerator Statement: The outcome being measured is acute, unplanned hospital visits (ED visit, observation stay, or unplanned inpatient admission) occurring within 7 days of a general surgery procedure performed at an ASC.
Denominator Statement: Target Population

Included patients:
The target population for this measure is Medicare FFS patients aged 65 years and older, who are undergoing outpatient general surgery procedures in ASCs that are within the scope of general surgery training. Specifically, the cohort of procedures includes the following types of surgeries: abdominal, alimentary tract, breast, skin/soft tissue, wound, and varicose vein.
The Medicare FFS population was chosen because of the availability of a national dataset (Medicare claims) that could be used to develop, test, and publicly report the measure. We limit the measure to patients who have been enrolled in Medicare FFS Parts A and B for the 12 months prior to the date of surgery to ensure that we have adequate data for identifying comorbidities for risk adjustment.

Included procedures:
The target group of procedures is surgical procedures that (1) are routinely performed at ASCs, (2) involve risk of post-surgery hospital visits, and (3) are within the scope of general surgery training. The scope of general surgery overlaps with that of other specialties (for example, vascular surgery and, plastic surgery). For this measure, we targeted surgeries that general surgeons are trained to perform with the understanding that other subspecialists may also be performing many of these surgeries at ASCs. Since the type of surgeon performing a particular procedure may vary across ASCs in ways that affect quality, the measure is neutral to surgeons' specialty training.

To identify eligible ASC general surgery procedures, we first identified a list of procedures from Medicare's 2014 and 2015 ASC lists of covered procedures, which include procedures for which ASCs can be reimbursed under the ASC payment system. This lists of surgeries is publicly available at: https://www.cms.gov/medicare/medicare-fee-for-service-payment/ascpayment/11_addenda_updates.html (download January 2014 and January 2015 ASC Approved HCPCS Code and Payment Rates, Addendum AA). Surgeries on the ASC list of covered procedures do not involve or require: major or prolonged invasion of body cavities, extensive blood loss, major blood vessels, or care that is either emergent or life-threatening. The ASC list is annually reviewed and updated by Medicare, and includes a transparent public comment submission and review process for addition and/or removal of procedure codes. Using an existing, defined list of surgeries, rather than defining surgeries de novo, is useful for long-term measure maintenance. Procedures listed in Medicare's list of covered ASC procedures are defined using Healthcare Common Procedure Coding System (HCPCS) and Common Procedural Terminology (CPT®) codes.

Ambulatory procedures include a heterogeneous mix of non-surgical procedures, minor surgeries, and more substantive surgeries. The measure is not intended to include very low-risk (minor) surgeries or non-surgical procedures, which typically have a high volume and a very low outcome rate. Therefore, to focus the measure only on the subset of surgeries on Medicare's list of covered ASC procedures that impose a meaningful risk of post-procedure hospital visits, the measure includes only "major" and "minor" procedures, as indicated by the Medicare Physician Fee Schedule global surgery indicator (GSI) values of 090 and 010, respectively. The GSI code reflects the number of post-operative days that are included in a given procedure's global surgical payment and identifies surgical procedures of greater complexity and follow-up care. This list of GSI values is publicly available for calendar year (CY) 2014 at: https://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/PhysicianFeeSched/PFS-Federal-Regulation-Notices-Items/CMS-1600-FC.html and for CY 2015 at: https://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/PhysicianFeeSched/PFS-Federal-Regulation-Notices-Items/CMS-1612-FC.html (download PFS Addenda, Addendum B).

Finally, to identify the subset of general surgery ASC procedures, we reviewed with consultants and Technical Expert Panel (TEP) members the Clinical Classifications Software (CCS) categories of procedures developed by the Agency for Healthcare Research and Quality (AHRQ). We identified and included CCS categories within the scope of general surgery, and only included individual procedures within the CCS categories at the procedure (CPT® code) level if they were within the scope of general surgery practice. We did not include in the measure gastrointestinal endoscopy, endocrine, or vascular procedures, other than varicose vein procedures, because reasons for hospital visits are typically related to patients' underlying comorbidities.

See the attached Data Dictionary, sheet S.9 "Codes Used to Define Cohort" for a complete list of all CPT procedure codes included in the measure cohort.

Denominator Exclusions: The measure excludes surgeries for patients without 7 or more days of continuous enrollment in Medicare FFS Parts A and B after the surgery. The measure excludes these patients to ensure all patients have full data available for outcome assessment.

Measure Type: Outcome
Data Source: Claims, Enrollment Data
Level of Analysis: Facility

IF Endorsement Maintenance – Original Endorsement Date: N/A Most Recent Endorsement Date: N/A

New Measure -- Preliminary Analysis

Criteria 1: Importance to Measure and Report

1a. Evidence

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.

Evidence Summary – This measure will identify ambulatory surgical centers (ASC) that have significantly higher rates of unplanned hospital visits related to other ASCs performing the same types of procedures on similar patients. In the literature, hospital visit rates following outpatient surgery vary from 0.5-9.0%. This measure is based on literature suggesting that patient selection and preparation, post-operative care, and post-discharge planning can affect the rate of adverse events and unplanned admissions following outpatient surgery.

Empirical data demonstrating a relationship between the outcome to at least one healthcare process is now required. NQF guidance states that a wide variation in performance can be used as evidence, assuming the data are from a robust number of providers and results are not subject to systematic bias.

Question for the Committee:

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

- *Is the performance data from the literature sufficient, in size and variance, to demonstrate that some ASC facilities are engaging in quality improvement activities to decrease unplanned hospital admissions after surgery better than others?*

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

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

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

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

- The developer assessed ASC-level variation in performance scores using Medicare FFS claims data for fiscal years 2014 and 2015, which included 236,999 general surgeries from 1,642 ASCs.
- The developer reports variation in the risk adjusted measure scores, ranging from 0.42 to 2.13.

Disparities

- The developer evaluated disparities with the observed rate and then evaluated the magnitude of association of three risk factors (dual eligible, race, SES) with the outcome after adjustment. Dual eligible, African Americans, and those with AHRQ SES Index scores below 42.7 had higher observed rates.
- The developer concluded that the risk factors have a modest but statistically significant association with the risk of a hospital visit.

Disparity Marker	Observed Rate (%)
Dual Eligible	3.7
Non dual eligible	2.2
African American	3.1
Non African American	2.2
AHRQ SES Index <42.7	2.7
AHRQ SES Index >42.7	2.2

Disparity Marker	Odds Ratio	Confidence Interval (95%)	*p* value
Dual Eligible	1.34	1.22-1.48	<0.0001
Race	1.23	1.06-1.42	0.005
AHRQ SES Index	1.14	1.06-1.22	0.0004

Questions for the Committee:

- *Is there a gap in unplanned hospital visits following ambulatory surgical visits that warrants a national performance measure?*
- *Are you aware of evidence that other disparities exist in this area of healthcare?*

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

Committee pre-evaluation comments

Criteria 1: Importance to Measure and Report (including 1a, 1b, 1c)

1a. Evidence to Support Measure Focus:

**Sufficient data are presented to judge the measure performance (i.e. Medicare claims data nationwide)
**Is there at least one thing that the provider can do to achieve ... Yes I believe there is both in the literature and in my experience as a surgeon working in ASCs
Is the performance data from the literature sufficient, ... yes I believe the evidence is solid.
**The project is very logical and in many ways follows the methodology of 30 day readmission measures that we have evaluated. This will capture unplanned in-patient hospital admissions that occur within 7 days of general surgical procedures that are performed in ambulatory surgery centers and will risk adjust based on administrative data and some socioeconomic and racial factors. Centers will be evaluated for their relative performance assuming that those

with lower rates of unplanned admissions are either having superior operative results with less complications and/or better processes of care in patient education and follow up communication than those that have higher rates, assuming very importantly that the patient's relative risk is indeed being accurately captured, which I have some concern about.
**Evidence well supported
**Evidence is sufficient
**Outcome measure with good data to support it

1b. Performance Gap

**Yes, substantially variability was demonstrated.
**Yes. While the performance gap is less than I expected, it is statistically significant. I believe that widespread reporting of and attention to this measure will likely improve quality of care in ASCs and consequently, patient outcomes.
I'm not aware of disparities in this area above and beyond those presented by the developers."
**This is a relatively new area being developed and an important one because surgical centers probably are not as closely scrutinized as hospitals leaving a considerable gap in evaluation of outcomes and processes, so this measure is needed to fill that gap. This should help to define the degree of gap, as I assume there probably is, and then to hopefully close that, since most surgeons respond to this type of comparative data. The fact that there is considerable variation in the incidence of 7 day admissions indicates that there is a need to close the gap between the centers.
**PG exists
**Performance gap is well-described
**Performance gap is provided and demonstrates a gap

Criteria 2: Scientific Acceptability of Measure Properties
2a. Reliability: Specifications and Testing
2b. Validity: Testing; Exclusions; Risk-Adjustment; Meaningful Differences; Comparability Missing Data

Reliability

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

2a2. Reliability testing demonstrates if the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise enough to distinguish differences in performance across providers. For maintenance measures – less emphasis if no new testing data provided.

Validity

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

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

Complex measure evaluated by Scientific Methods Panel? ☒ **Yes** ☐ **No**
Evaluators: Sherrie Kaplan, Christie Tiegland, Laurent Glance

Evaluation of Reliability and Validity:

Evaluation A
Evaluation B
Evaluation C

Questions for the Committee regarding reliability:

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

Questions for the Committee regarding validity:

- *Do you have any concerns regarding the validity of the measure (e.g., exclusions, risk-adjustment approach, etc.)?*
- *The Scientific Methods Panel is satisfied with the validity analyses for the measure. Does the Committee think there is a need to discuss and/or vote on validity?*

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

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

Committee pre-evaluation comments

Criteria 2: Scientific Acceptability of Measure Properties (including all 2a, 2b, and 2c)

2a1. Reliability-Specifications

**No issues.

**Evaluation A raises some excellent points regarding the limitations of the disparities analysis (i.e. limiting race to two categories, analyzing SES at the zip code level etc.). However, I find the rationale provided in the measure submission adequately responds to these concerns.

I do believe the measure could be successfully implemented and I have no major concerns with the reliability or validity testing.

I do not see a need to discuss or vote on validity

**I believe the developers have made a reasonably good case for the reproducibility and reliability of collecting this data.

**Reliable

**Well-defined

All clearly defined

**Despite the reply from the stewards, the C-stasitics offered are on the margin. Overall community effect of the population of the ASC and its effect on performance within the measure across the the three given SES parameters is not clear, and its dismissal is not as well. The stewards might want to consider the incorporation of some of the specific individual ICD 9 (now 10) codes that were brought into the Risk Stratified Episode of Care Cost Measure for THA/TKA that CORE developed previously, especially the neuro-degenerative/neuro-cognitive codes and the more specific codes re: obesity. The patients with higher HCC risk factors might be over populating the return to hospital statistics because of returns unrelated to the surgery; perhaps a the longitudinal rate of hospital encounters pretending the index event could be used for a separate risk factor?

2a2. Reliability - Testing

**No

**no

**I would consider the reliability at least as moderate, being concerned somewhat about the accuracy of administrative data which is generally not audited to the extent that clinical databases audit.

**No

**No concerns

**No

**The concern is the risk adjustment, especially if used across small populations and small percentile differences in a payment program.

2b1. Validity

2b4-7. Threats to Validity

2b4. Meaningful Differences

**No issues

**no

**I have some concerns about the validity related to the risk adjustment process, both social and co-morbidities, etc. I realize that there are several papers from centers and individuals that I respect comparing administrative to chart abstracted data but I still have concerns about the degree of severity of various co-morbidities or the lack of that data with the administrative data. Exclusions do not seem to be an issue in this protocol.

**Valid

**No

**Not a substantial threat

**This measure borrows on previous validation work regarding the validity of the administrative data set and real chart review. It is not clear that that conclusion can be assumed.

2b2-3. Other Threats to Validity (Exclusions, Risk Adjustment)
**Minimal risk adjustment (other than socioeconomic) applied.
**no
**In regard to the social risks, zip codes have short comings because a zip code area can have a mixture of socioeconomic neighborhoods, and the racial diversity of the US is not captured in the data presented. I probably would rate the validity as moderate. It is somewhat reassuring that the C-index is 0.69, not great, but reasonably good.
Risk adjusted via admin DB
**There are substantial differences among AA, dual-eligible and low SES populations which CMS plans to adjust for.
**Reasonable risk adjustment as much as possible with a measure specified in claims
**Please see comments under reliability. Concerns exist, especially if the measure is used to adjust payments based on small differences in percentile performance.

Criterion 3. Feasibility

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.

- This is a claims based measure. No data elements are in electronic sources.
- There are no fees, licensing, or other requirements to use this measure as specified.

Questions for the Committee:
- *Are the required data elements routinely generated and used during care delivery?*
- *Is the data collection strategy ready to be put into operational use?*

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

Committee pre-evaluation comments
Criteria 3: Feasibility

3. Feasibility
**Can be captured in national CMS billing data and calculated remotely.
**I agree with NQF staff this this is high.
**I believe that most of the proposed data points are reasonably straight forward and that the project is very feasible.
**Feasible
**No concerns - this measure can be specified in claims
**No concerns
** Please see comments under reliability. Concerns exist, especially if the measure is used to adjust payments based on small differences in percentile performance.

Criterion 4: Usability and Use

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

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

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

Current uses of the measure
Publicly reported? ☐ **Yes** ☒ **No**

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

OR

Planned use in an accountability program? ☒ Yes ☐ No

Accountability program details

- The developer reports that this measure may ultimately be used in one or more Centers for Medicare & Medicaid Services' (CMS) programs, such as the Ambulatory Surgical Center Quality Reporting Program (ASCQR).
    - This measure was approved for consideration under the ASCQR program and was discussed by the MAP Hospital Workgroup in December 2017. MAP conditionally supported this measure for the ASCQR program pending NQF review and endorsement. MAP recognized that this measure assesses an important outcome for patients receiving care at ambulatory surgery centers and addresses crucial safety concerns by tracking if a patient requires treatment at an acute care hospital (including emergency department (ED) visits, observation stays, and unplanned inpatient admissions) within 7 days of the procedure performed at an ASC. MAP noted this measure could help balance incentives to perform more procedures on an outpatient basis. However, MAP acknowledged a number of concerns raised in public comments about the measure. Commenters raised concerns about the attribution model of measure, noting that these are relatively rare events and could disproportionately impact low-volume ASCs, and that the measure may need risk adjustment for social risk factors. MAP noted this measure should be submitted for NQF endorsement to assess the potential impact of these concerns on the reliability and validity of the measure.

4a.2. Feedback on the measure by those being measured or others. Three criteria demonstrate feedback: 1) those being measured have been given performance results or data, as well as assistance with interpreting the measure results and data; 2) those being measured and other users have been given an opportunity to provide feedback on the measure performance or implementation; 3) this feedback has been considered when changes are incorporated into the measure

Feedback on the measure by those being measured or others

- The developer reports that they recruited a national TEP during measure development and hosted a public comment period. TEP members and commenters included representatives of the ASCs.
- Data and results were provided to the TEP and members of the TEP were able to give input on five occasions during the measure development process.
- Revisions made to the measure based on feedback included: renaming the measure to reflect the procedures included in the measure cohort; removal of 15 individual CPT codes that were outside the scope of general surgery practice; and a review of variables for the final risk model where one was retained (opioid use) since experts felt it was an important risk predictor.

Additional Feedback: Not applicable

***Questions for the Committee**:*

- *How 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 (4a1. Improvement; 4a2. Benefits of measure)

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

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

Improvement results

- The developer indicated that the question was not applicable since the measure is not yet in use.

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 indicated that the question was not applicable since the measure is not yet in use.

Potential harms:

- The developer indicated that the question was not applicable since the measure is not yet in use.

Additional Feedback:

- The developer indicated that the question was not applicable since the measure is not yet in use.

***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

Criteria 4: Usability and Use

4a1. Use - Accountability and Transparency

**Not yet in use

**The committee asked how measure can be used to further the goal of high quality healthcare? This is a ubiquitous wide-lens view of ASC performance. A focus on reducing readmissions is likely to increase attention to detail in the pre, intra, and post op care of ASC patients. With a national trend towards pushing procedures more into the outpatient realm, this measure can be an important check to be sure we are not pushing too hard to the detriment of patients.

**There is significant variation in rates of 7 day admission across the various surgery centers which allows the centers to identify opportunities for improvement after seeing the data analysis.

**Usable

**New measure - but MAP approved it for use going forward

**Not currently

**Please see comments under reliability. Concerns exist, especially if the measure is used to adjust payments based on small differences in percentile performance.

4b1. Usability – Improvement

**This measure should be a good one to drive performance.

**I do not understand staff comments in this area

**In general, responsible surgeons and other clinicians want to perform to the best of their abilities and if data shows them to not be performing well this should stimulate them to try to understand why and make some changes to improve the outcomes. This has worked well in other settings, but only if they have confidence in the data and its analysis.

**No concerns

**Unknown at this point. IF the risk adjustment is insufficient to account for the substantial differences in outcome among low SES populations, public reporting of this measure could result in poor access to outpatient care in that population

**Benefits likely outweigh harms

**The measure would be best used first in public reporting so that potential deficits in risk adjustment and SES risk adjustment could be assessed before moving to payment adjustments.

Criterion 5: Related and Competing Measures

Related or competing measures

- 2539 Facility 7-Day Risk-Standardized Hospital Visit Rate after Outpatient Colonoscopy
- 2687 Hospital Visits after Hospital Outpatient Surgery
- 3366 Hospital Visits after Urology Ambulatory Surgical Center Procedures (currently under consideration by the Surgery Standing Committee)

Harmonization

- The developer reports that the measure specification are harmonized with the above listed related measures.

Committee pre-evaluation comments

Criterion 5: Related and Competing Measures

Public and member comments

Comments and Member Support/Non-Support Submitted as of: January 9, 2018

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

Evaluation A

Scientific Acceptability

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

Instructions:

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

Measure Number: 3357
Measure Title: Facility-Level 7-Day Hospital Visits after General Surgery Procedures Performed at Ambulatory Surgical Centers

RELIABILITY

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

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

   ☐Yes (go to Question #2)

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

   The time period for measurement is not entirely clear to me. They indicate they are using 2 years of data consistent with CMS new practices, but it seems that the initial 12 months is required to gather the risk

factor variables and that any procedures during first 12 months will not be included in measure because those patients would not have prior 12 months of data? And the period ends at least 7 days before the end of the measurement year? I think that is what they are saying but it is not clearly stated as to what the exact measurement period is.

Also, I do not see any exclusions from reporting for ASCs that may have a minimal number of the type of procedures included in the measure and rates may thus be unstable or not true indicator of quality for that ASC. The algorithm did control for differences in numbers of procedures but not sure that is sufficient for small denominators.

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

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

   ☒Yes (go to Question #4)

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

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

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

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

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

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

   ☒Yes (go to Question #5)

   ☐No (go to Question #8)

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

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

   ☒Yes (go to Question #6)

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

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

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

   ☐High (go to Question #8)

   ☒Moderate (go to Question #8)

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

7. Was other reliability testing reported?
   ☐Yes (go to Question #8)
   ☐No (rate Question #11: OVERALL RELIABILITY as LOW and proceed to the VALIDITY SECTION)

8. Was reliability testing conducted with patient-level data elements that are used to construct the performance measure?
   TIPS: Prior reliability studies of the same data elements may be submitted; if comparing abstraction to "authoritative source/gold standard" see Validity Section Question #15)
   ☒Yes (go to Question #9)
   ☐No (if there is score-level testing, rate Question #11: OVERALL RELIABILITY based on score-level rating from Question #6; otherwise, rate Question #11: OVERALL RELIABILITY as INSUFFICIENT. Then proceed to the VALIDITY SECTION)

9. Was the method described and appropriate for assessing the reliability of ALL critical data elements?
   TIPS: For example: inter-abstractor agreement (ICC, Kappa); other accepted method with description of how it assesses reliability of the data elements
   Answer no if: only assessed percent agreement; did not assess separately for all data elements (at least numerator, denominator, exclusions)
   ☒Yes (go to Question #10)
   ☐No (if no, please explain below and rate Question #10 as INSUFFICIENT)

10. **RATING (data element)** – Based on the reliability statistic and scope of testing (number and representativeness of patients and entities), what is the level of certainty or confidence that the data used in the measure are reliable?
    TIPS: Consider the following: Is the test sample adequate to generalize for widespread implementation? Can data elements be collected consistently?
    ☒Moderate (if score-level testing was NOT conducted, rate Question #11: OVERALL RELIABILITY as MODERATE)
    ☐Low (if score-level testing was NOT conducted, rate Question #11: OVERALL RELIABILITY as LOW)
    ☐Insufficient (go to Question #11)

11. OVERALL RELIABILITY RATING

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

☐High (NOTE: Can be HIGH only if score-level testing has been conducted)
☒Moderate (NOTE: Moderate is the highest eligible rating if score-level testing has not been conducted)
☐Low (please explain below) [NOTE: Should rate LOW if you believe specifications are NOT precise, unambiguous, and complete]
☐Insufficient (please explain below) [NOTE: For most measure types, testing at both the score level and the data element level is not required]

VALIDITY

Assessment of Threats to Validity

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

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

   ☐Yes (go to Question #2)

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

Empirical testing for data element validity did not cover ALL critical data elements in my mind, specifically the SES proxy used which is measured using a very small survey sample in most geographic areas and aggregated at ZIP code level which can cover widely disparate populations in many geographic ZIPs thus averaging out the social risk factors and resulting in little/no impact on the outcome.

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

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

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

   ☒No (go to Question #3)

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

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

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

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

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

   c. Any concerns regarding the risk-adjustment approach?

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

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

   ☐No (go to Question #4)

   The social risk factor data elements did not have sufficient reliability testing at the patient level. Though a validated SES composite score was used, it was calculated using ACS block level data (a very small sample) at the 5 digit ZIP level, which comprises a wide population that can have widely varying SES within the ZIP area, resulting in "averaging out" and thus showing little impact overall. In addition, race/ethnicity was define as African American vs. Other, thus Hispanics, Asians and other race/ethnic groups are lumped in with "White" which also can skew any impact of race/ethnicity. Given they found effects of SES using the crude ZIP level survey sample data, using social risk factor data collected at a more granular level could very likely show more significant differences that would also impact the outcome rates significantly after controlling for other risk factors included in the models.

I also have concerns about using HCCs as the data level for chronic conditions. Certain individual conditions within an HCC are often more highly associated with the outcome but that relationship gets lost the hierarchical category. An HCC may not be highly associated but individual conditions may be very highly associated. The HCCs also do not necessarily capture the impact of having multiple conditions that may be combined in one HCC.

4. Analysis of potential threats to validity: Any concerns regarding ability to identify meaningful differences in performance or overall poor performance?
   ☒Yes (please explain below then go to Question #5)
   ☐No (go to Question #5)

Concerns about the percentages shown for the quartile cut-offs and how well they truly capture the intent of the variable at the ASC level. For example, the 1st quartile cut-off for the proportion of Medicaid dual eligible patients at the ASC level is <=1.82% and for the 4th quartile >= 7.06%. 7% is still a VERY LOW proportion of dual eligible patients, indicating the distribution of dual eligible patients having one of the outpatient surgeries at an ASC seems to be unexpectedly low at almost all the ASCs included in the sample. The 4th quartile includes ASCs with only 7% duals up to ASCs with 100% duals potentially, which could be why we didn't see this contributing to ASC level differences in rates when including this as a social risk factor adjuster.

5. Analysis of potential threats to validity: Any concerns regarding comparability of results if multiple data sources or methods are specified?
   ☐Yes (please explain below then go to Question #6)
   ☐No (go to Question #6)
   ☒Not applicable (go to Question #6)

6. Analysis of potential threats to validity: Any concerns regarding missing data?
   ☐Yes (please explain below then go to Question #7)
   ☒No (go to Question #7)

Assessment of Measure Testing

7. Was <u>empirical</u> validity testing conducted using the measure as specified and appropriate statistical test?
   Answer no if: face validity; only refer to clinical evidence; only descriptive statistics; only describe process for data management/cleaning/computer programming; testing does not match measure specifications (i.e. data, eMeasure, level, setting, patients).
   ☐Yes (go to Question #10) [NOTE: If appropriate empirical testing has been conducted, then evaluation of face validity is not necessary. Go to Question #8 **only if** there is insufficient information provided to evaluate data element and score-level testing.]
   ☒No (please explain below then go to Question #8)
   Face validity only (however, they refer to prior empirical validity testing "For several other NQF-endorsed measures, our team has demonstrated the validity of using claims data for risk adjustment in lieu of medical record data in estimating facility-level measure scores.")

8. Was <u>face validity</u> systematically assessed by recognized experts to determine agreement on whether the computed performance measure score from the measure as specified can be used to distinguish good and poor quality?
   TIPS: Answer no if: focused on data element accuracy/availability/feasibility/other topics; the degree of consensus and any areas of disagreement not provided/discussed.

☒Yes (go to Question #9)
☐No (please explain below and rate Question #17: OVERALL VALIDITY as INSUFFICIENT)

9. **RATING (face validity)** - Do the face validity testing results indicate substantial agreement that the performance measure score from the measure as specified can be used to distinguish quality AND potential threats to validity are not a problem, OR are adequately addressed so results are not biased?
   ☒Yes (if a NEW measure, rate Question #17: OVERALL VALIDITY as MODERATE)
   ☐ Yes (if a MAINTENANCE measure, do you agree with the justification for not conducting empirical testing? If no, rate Question #17: OVERALL VALIDITY as INSUFFICIENT; otherwise, rate Question #17: OVERALL VALIDITY as MODERATE)
   ☐No (please explain below and rate Question #17: OVERALL VALIDITY AS LOW)

10. Was validity testing conducted with computed performance measure scores for each measured entity?
    TIPS: Answer no if: one overall score for all patients in sample used for testing patient-level data.
    ☐Yes (go to Question #11)
    ☐No (please explain below and go to Question #13)

11. Was the method described and appropriate for assessing conceptually and theoretically sound hypothesized relationships?
    TIPS: For example: correlation of the performance measure score on this measure and other performance measures; differences in performance scores between groups known to differ on quality; other accepted method with description of how it assesses validity of the performance score
    ☐Yes (go to Question #12)
    ☐No (please explain below, rate Question #12 as INSUFFICIENT and then go to Question #14)

12. **RATING (measure score)** - Based on the measure score results (significance, strength) and scope of testing (number of measured entities and representativeness) and analysis of potential threats, what is the level of certainty or confidence that the performance measure scores are a valid indicator of quality?
    ☐High (go to Question #14)
    ☐Moderate (go to Question #14)
    ☐Low (please explain below then go to Question #13)
    ☐Insufficient

13. Was other validity testing reported?
    ☐Yes (go to Question #14)
    ☐No (please explain below and rate Question #17: OVERALL VALIDITY as LOW)

14. Was validity testing conducted with patient-level data elements?
    TIPS: Prior validity studies of the same data elements may be submitted
    ☐Yes (go to Question #15)
    ☐No (please explain below and rate Question #17: OVERALL VALIDITY as INSUFFICIENT if no score-level testing was conducted, otherwise, rate Question #17: OVERALL VALIDITY based on score-level rating from Question #12)

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

TIPS: For example: Data validity/accuracy as compared to authoritative source- sensitivity, specificity, PPV, NPV; other accepted method with description of how it assesses validity of the data elements.
Answer No if: only assessed percent agreement; did not assess separately for all data elements (at least numerator, denominator, exclusions)

☐Yes (go to Question #16)

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

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

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

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

☐Insufficient (go to Question #17)

17. OVERALL VALIDITY RATING

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

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

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

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

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

See comments above re SES data used and use of HCCs; feel the issues are strong enough to require further empirical validation. Would like to see results further stratified by percent dual population in ASC for example, comparing not all ASCs with 7% or more of patients served having dual status but rates for ASCs with 80% or more of population served having dual status. I am unconvinced the data used for race/ethnicity and SES is granular and accurate enough to actually capture the impact of those risk factors on the outcome.

FOR COMPOSITE MEASURES ONLY: Empirical analyses to support composite construction

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

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

☐High

☐Moderate

☐Low (please explain below)

☐Insufficient (please explain below)

Evaluation B

Scientific Acceptability

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

Instructions:

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

Measure Number: 3357
Measure Title:
Facility-Level 7-Day Hospital Visits after General Surgery Procedures Performed at Ambulatory Surgical Centers

RELIABILITY

11. Are submitted specifications precise, unambiguous, and complete so that they can be consistently implemented? *NOTE: NQF staff will conduct a separate, more technical, check of eMeasure (eCQM) specifications, value sets, logic, and feasibility, so no need to consider these in your evaluation.*
    TIPS: Consider the following: Are all the data elements clearly defined? Are all appropriate codes included? Is the logic or calculation algorithm clear? Is it likely this measure can be consistently implemented?
    ☒Yes (go to Question #2)
    ☐No (please explain below, and go to Question #2) NOTE that even though ***non-precise specifications should result in an overall LOW rating for reliability***, we still want you to look at the testing results.

12. Was empirical reliability testing (at the data element or measure score level) conducted using statistical tests with the measure as specified?
    TIPS: Check the 2nd "NO" box below if: only descriptive statistics provided; only describes process for data management/cleaning/computer programming; testing does not match measure specifications (i.e. data, eMeasure, level of analysis, patients)
    ☒Yes (go to Question #4)
    ☐No, there is reliability testing information, but *not* using statistical tests and/or not for the

measure as specified OR there is no reliability testing (please explain below then go to Question #3)

13. Was **empirical VALIDITY testing** of patient-level data conducted?
    ☐Yes (use your rating from data element validity testing – Question #16- under Validity Section)
    ☐No (please explain below and rate Question #11: OVERALL RELIABILITY as INSUFFICIENT and proceed to the VALIDITY SECTION)

14. Was reliability testing conducted with computed performance measure scores for each measured entity?
    TIPS: Answer no if: only one overall score for all patients in sample used for testing patient-level data
    ☒Yes (go to Question #5)
    ☐No (go to Question #8)

15. Was the method described and appropriate for assessing the proportion of variability due to real differences among measured entities? *NOTE: If multiple methods used, at least one must be appropriate.*
    TIPS: Examples of appropriate methods include signal-to-noise analysis (e.g. Adams/RAND tutorial); random split-half correlation; other accepted method with description of how it assesses reliability of the performance score.
    ☒Yes (go to Question #6)
    ☐No (please explain below then go to Question #8)

    The ICC – used to examine measure reliability – was 0.51. Values less than 0.5 are indicative of poor agreement, and values between 0.5 and 0.75 are consistent with moderate agreement. A value of 051 is right at the margin – and hence more consistent with poor-to-moderate agreement.

16. **RATING (score level)** - What is the level of certainty or confidence that the performance measure scores are reliable?
    TIPS: Consider the following: Is the test sample adequate to generalize for widespread implementation? Do the results demonstrate sufficient reliability so that differences in performance can be identified?
    ☐High (go to Question #8)
    ☒Moderate (go to Question #8)
    ☐Low (please explain below then go to Question #7)

    The ICC – used to examine measure reliability – was 0.51. Values less than 0.5 are indicative of poor agreement, and values between 0.5 and 0.75 are consistent with moderate agreement. A value of 051 is right at the margin – and hence more consistent with poor-to-moderate agreement.

17. Was other reliability testing reported?
    ☐Yes (go to Question #8)
    ☐No (rate Question #11: OVERALL RELIABILITY as LOW and proceed to the VALIDITY SECTION)

18. Was reliability testing conducted with patient-level data elements that are used to construct the performance measure?
    TIPS: Prior reliability studies of the same data elements may be submitted; if comparing abstraction to "authoritative source/gold standard" see Validity Section Question #15)
    ☒Yes (go to Question #9)

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

19. Was the method described and appropriate for assessing the reliability of ALL critical data elements?
    TIPS: For example: inter-abstractor agreement (ICC, Kappa); other accepted method with description of how it assesses reliability of the data elements
    Answer no if: only assessed percent agreement; did not assess separately for all data elements (at least numerator, denominator, exclusions)
    ☒Yes (go to Question #10)
    ☐No (if no, please explain below and rate Question #10 as INSUFFICIENT)

20. **RATING (data element)** – Based on the reliability statistic and scope of testing (number and representativeness of patients and entities), what is the level of certainty or confidence that the data used in the measure are reliable?
    TIPS: Consider the following: Is the test sample adequate to generalize for widespread implementation? Can data elements be collected consistently?
    ☒Moderate (if score-level testing was NOT conducted, rate Question #11: OVERALL RELIABILITY as MODERATE)
    ☐Low (if score-level testing was NOT conducted, rate Question #11: OVERALL RELIABILITY as LOW)
    ☐Insufficient (go to Question #11)

11. OVERALL RELIABILITY RATING

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

☐High (NOTE: Can be HIGH only if score-level testing has been conducted)
☒Moderate (NOTE: Moderate is the highest eligible rating if score-level testing has not been conducted)
☐Low (please explain below) [NOTE: Should rate LOW if you believe specifications are NOT precise, unambiguous, and complete]
☐Insufficient (please explain below) [NOTE: For most measure types, testing at both the score level and the data element level is not required]

VALIDITY

Assessment of Threats to Validity

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

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

☒Yes (go to Question #2)

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

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

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

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

☒No (go to Question #3)

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

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

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

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

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

c. Any concerns regarding the risk-adjustment approach?

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

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

☒No (go to Question #4)

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

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

☒No (go to Question #5)

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

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

☒No (go to Question #6)

☐Not applicable (go to Question #6)

22. Analysis of potential threats to validity: Any concerns regarding missing data?
☐Yes (please explain below then go to Question #7)
☒No (go to Question #7)

Assessment of Measure Testing

23. Was empirical validity testing conducted using the measure as specified and appropriate statistical test?
Answer no if: face validity; only refer to clinical evidence; only descriptive statistics; only describe process for data management/cleaning/computer programming; testing does not match measure specifications (i.e. data, eMeasure, level, setting, patients).
☐Yes (go to Question #10) [NOTE: If appropriate empirical testing has been conducted, then evaluation of face validity is not necessary. Go to Question #8 **only if** there is insufficient information provided to evaluate data element and score-level testing.]
☒No (please explain below then go to Question #8)

24. Was face validity systematically assessed by recognized experts to determine agreement on whether the computed performance measure score from the measure as specified can be used to distinguish good and poor quality?
TIPS: Answer no if: focused on data element accuracy/availability/feasibility/other topics; the degree of consensus and any areas of disagreement not provided/discussed.
☒Yes (go to Question #9)
☐No (please explain below and rate Question #17: OVERALL VALIDITY as INSUFFICIENT)

25. **RATING (face validity)** - Do the face validity testing results indicate substantial agreement that the performance measure score from the measure as specified can be used to distinguish quality AND potential threats to validity are not a problem, OR are adequately addressed so results are not biased?
☒Yes (if a NEW measure, rate Question #17: OVERALL VALIDITY as MODERATE)
☐ Yes (if a MAINTENANCE measure, do you agree with the justification for not conducting empirical testing? If no, rate Question #17: OVERALL VALIDITY as INSUFFICIENT; otherwise, rate Question #17: OVERALL VALIDITY as MODERATE)
☐No (please explain below and rate Question #17: OVERALL VALIDITY AS LOW)

26. Was validity testing conducted with computed performance measure scores for each measured entity?
TIPS: Answer no if: one overall score for all patients in sample used for testing patient-level data.
☐Yes (go to Question #11)
☐No (please explain below and go to Question #13)

27. Was the method described and appropriate for assessing conceptually and theoretically sound hypothesized relationships?
TIPS: For example: correlation of the performance measure score on this measure and other performance measures; differences in performance scores between groups known to differ on quality; other accepted method with description of how it assesses validity of the performance score
☐Yes (go to Question #12)
☐No (please explain below, rate Question #12 as INSUFFICIENT and then go to Question #14)

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

☐High (go to Question #14)
☐Moderate (go to Question #14)
☐Low (please explain below then go to Question #13)
☐Insufficient

29. Was other validity testing reported?

☐Yes (go to Question #14)
☐No (please explain below and rate Question #17: OVERALL VALIDITY as LOW)

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

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

☐Yes (go to Question #15)
☐No (please explain below and rate Question #17: OVERALL VALIDITY as INSUFFICIENT if no score-level testing was conducted, otherwise, rate Question #17: OVERALL VALIDITY based on score-level rating from Question #12)

The measure developers did not specifically test the validity of patient-data elements for this specific measure. As per the measure developers, "While the applicability of these findings to our measure may be limited because these medical record validations medical record evaluations were focused on patients admitted for specific medical conditions, they nevertheless suggest claims data generally have an acceptable degree of agreement with clinical data at a facility level." However, since data element validation on prior measures is also based on a look-back period of 12 months, I believe that prior validation of data elements is generally applicable to this measure.

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

TIPS: For example: Data validity/accuracy as compared to authoritative source- sensitivity, specificity, PPV, NPV; other accepted method with description of how it assesses validity of the data elements.
Answer No if: only assessed percent agreement; did not assess separately for all data elements (at least numerator, denominator, exclusions)

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

I rated this as a "yes" – even though the measure developers do not report the validation results for this specific measure. But I think that the measure developers should summarize the results of data validation for the previous NQF-endorse readmission measure since both metrics capture similar outcomes.

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

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

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

17. OVERALL VALIDITY RATING

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

☐High (NOTE: Can be HIGH only if score-level testing has been conducted)
☒Moderate (NOTE: Moderate is the highest eligible rating if score-level testing has NOT been conducted)
☐Low (please explain below) [NOTE: Should rate LOW if you believe that there are threats to validity and/or threats to validity were not assessed]
☐Insufficient (if insufficient, please explain below) [NOTE: For most measure types, testing at both the score level and the data element level is not required] [NOTE: If rating is INSUFFICIENT for all empirical testing, then go back to Question #8 and evaluate any face validity that was conducted, then reconsider this overall rating.]

I scored this as "moderate" because measure reliability was empirically assessed as moderate.

FOR COMPOSITE MEASURES ONLY: Empirical analyses to support composite construction

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

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

☐High
☐Moderate
☐Low (please explain below)
☐Insufficient (please explain below)

Evaluation C

Scientific Acceptability

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

Instructions:

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

Measure Number: 3357
Measure Title: Facility-Level 7-Day Hospital Visits after General Surgery Procedures Performed at Ambulatory Surgical Centers

RELIABILITY

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

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

    ☒Yes (go to Question #2)

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

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

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

    ☒Yes (go to Question #4)

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

Question #3)

23. Was **empirical VALIDITY testing** of patient-level data conducted?
    - ☐Yes (use your rating from data element validity testing – Question #16- under Validity Section)
    - ☐No (please explain below and rate Question #11: OVERALL RELIABILITY as INSUFFICIENT and proceed to the VALIDITY SECTION)

24. Was reliability testing conducted with computed performance measure scores for each measured entity?
    TIPS: Answer no if: only one overall score for all patients in sample used for testing patient-level data
    - ☒Yes (go to Question #5)
    - ☐No (go to Question #8)

25. Was the method described and appropriate for assessing the proportion of variability due to real differences among measured entities? *NOTE: If multiple methods used, at least one must be appropriate.*
    TIPS: Examples of appropriate methods include signal-to-noise analysis (e.g. Adams/RAND tutorial); random split-half correlation; other accepted method with description of how it assesses reliability of the performance score.
    - ☒Yes (go to Question #6)

    But they used only split-half reliability. What is really needed is the ICC for between vs. within variance by facility, not agreement between samples.
    - ☐No (please explain below then go to Question #8)

26. **RATING (score level) -** What is the level of certainty or confidence that the performance measure scores are reliable?
    TIPS: Consider the following: Is the test sample adequate to generalize for widespread implementation?
    Do the results demonstrate sufficient reliability so that differences in performance can be identified?
    - ☐High (go to Question #8)
    - ☒Moderate (go to Question #8)
    - ☐Low (please explain below then go to Question #7)

27. Was other reliability testing reported?
    - ☐Yes (go to Question #8)
    - ☐No (rate Question #11: OVERALL RELIABILITY as LOW and proceed to the VALIDITY SECTION)

28. Was reliability testing conducted with patient-level data elements that are used to construct the performance measure?
    TIPS: Prior reliability studies of the same data elements may be submitted; if comparing abstraction to "authoritative source/gold standard" see Validity Section Question #15)
    - ☒Yes (go to Question #9)
    - ☐No (if there is score-level testing, rate Question #11: OVERALL RELIABILITY based on score-level rating from Question #6; otherwise, rate Question #11: OVERALL RELIABILITY as INSUFFICIENT. Then proceed to the VALIDITY SECTION)

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

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

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

☒Yes (go to Question #10)

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

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

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

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

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

☐Insufficient (go to Question #11)

11. OVERALL RELIABILITY RATING

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

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

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

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

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

VALIDITY

Assessment of Threats to Validity

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

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

☒Yes (go to Question #2)

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

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

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

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

☒No (go to Question #3)

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

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

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

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

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

c. Any concerns regarding the risk-adjustment approach?

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

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

☒No (go to Question #4)

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

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

☒No (go to Question #5)

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

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

☐No (go to Question #6)
☒Not applicable (go to Question #6)

38. Analysis of potential threats to validity: Any concerns regarding missing data?
☐Yes (please explain below then go to Question #7)
☒No (go to Question #7)

Assessment of Measure Testing

39. Was empirical validity testing conducted using the measure as specified and appropriate statistical test?
Answer no if: face validity; only refer to clinical evidence; only descriptive statistics; only describe process for data management/cleaning/computer programming; testing does not match measure specifications (i.e. data, eMeasure, level, setting, patients).
☐Yes (go to Question #10) [NOTE: If appropriate empirical testing has been conducted, then evaluation of face validity is not necessary. Go to Question #8 **only if** there is insufficient information provided to evaluate data element and score-level testing.]
☒No (please explain below then go to Question #8)
Only face validity was assessed

40. Was face validity systematically assessed by recognized experts to determine agreement on whether the computed performance measure score from the measure as specified can be used to distinguish good and poor quality?
TIPS: Answer no if: focused on data element accuracy/availability/feasibility/other topics; the degree of consensus and any areas of disagreement not provided/discussed.
☒Yes (go to Question #9)
☐No (please explain below and rate Question #17: OVERALL VALIDITY as INSUFFICIENT)

41. **RATING (face validity)** - Do the face validity testing results indicate substantial agreement that the performance measure score from the measure as specified can be used to distinguish quality AND potential threats to validity are not a problem, OR are adequately addressed so results are not biased?
☒Yes (if a NEW measure, rate Question #17: OVERALL VALIDITY as MODERATE)
☐ Yes (if a MAINTENANCE measure, do you agree with the justification for not conducting empirical testing? If no, rate Question #17: OVERALL VALIDITY as INSUFFICIENT; otherwise, rate Question #17: OVERALL VALIDITY as MODERATE)
☐No (please explain below and rate Question #17: OVERALL VALIDITY AS LOW)

42. Was validity testing conducted with computed performance measure scores for each measured entity?
TIPS: Answer no if: one overall score for all patients in sample used for testing patient-level data.
☐Yes (go to Question #11)
☐No (please explain below and go to Question #13)

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

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

☐Yes (go to Question #12)

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

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

☐High (go to Question #14)

☐Moderate (go to Question #14)

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

☐Insufficient

45. Was other validity testing reported?

☐Yes (go to Question #14)

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

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

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

☐Yes (go to Question #15)

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

Face validity was assessed as was discriminate validity (for disparities)

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

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

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

☐Yes (go to Question #16)

Using current NQF standards

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

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

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

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

☐Insufficient (go to Question #17)

17. OVERALL VALIDITY RATING

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

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

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

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

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

FOR COMPOSITE MEASURES ONLY: Empirical analyses to support composite construction

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

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

☐High

☐Moderate

☐Low (please explain below)

☐Insufficient (please explain below)

NATIONAL QUALITY FORUM—Evidence (subcriterion 1a)

Measure Number (*if previously endorsed*)**:** Click here to enter NQF number
Measure Title: Facility-Level 7-Day Hospital Visits after General Surgery Procedures Performed at Ambulatory Surgical Centers
IF the measure is a component in a composite performance measure, provide the title of the Composite Measure here: Click here to enter composite measure #/ title
Date of Submission: Click here to enter a date

Instructions

- *Complete 1a.1 and 1a.2 for all measures. If instrument-based measure, complete 1a.3.*
- *Complete **EITHER 1a.2, 1a.3 or 1a.4** as applicable for the type of measure and evidence.*
- *For composite performance measures:*
  - *A separate evidence form is required for each component measure unless several components were studied together.*
  - *If a component measure is submitted as an individual performance measure, attach the evidence form to the individual measure submission.*
- All information needed to demonstrate meeting the evidence subcriterion (1a) must be in this form. An appendix of *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.

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

1a. Evidence to Support the Measure Focus

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

- Outcome: [3] Empirical data demonstrate a relationship between the outcome and at least one healthcare structure, process, intervention, or service. If not available, wide variation in performance can be used as evidence, assuming the data are from a robust number of providers and results are not subject to systematic bias.
- Intermediate clinical outcome: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured intermediate clinical outcome leads to a desired health outcome.
- Process: [5] a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured process leads to a desired health outcome.
- Structure: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured structure leads to a desired health outcome.
- Efficiency: [6] evidence not required for the resource use component.
- For measures derived from patient reports, evidence should demonstrate that the target population values the measured outcome, process, or structure and finds it meaningful.
- Process measures incorporating Appropriate Use Criteria: See NQF's guidance for evidence for measures, in general; guidance for measures specifically based on clinical practice guidelines apply as well.

Notes

3. Generally, rare event outcomes do not provide adequate information for improvement or discrimination; however, serious reportable events that are compared to zero are appropriate outcomes for public reporting and quality improvement.
4. The preferred systems for grading the evidence are the Grading of Recommendations, Assessment, Development and Evaluation (GRADE) guidelines and/or modified GRADE.
5. Clinical care processes typically include multiple steps: assess → identify problem/potential problem → choose/plan intervention (with patient input) → provide intervention → evaluate impact on health status. If the measure focus is one step in such a multistep process, the step with the strongest evidence for the link to the desired outcome should be selected as the focus of measurement. Note: A measure focused only on collecting PROM data is not a PRO-PM.
6. Measures of efficiency combine the concepts of resource use and quality (see NQF's Measurement Framework: Evaluating Efficiency Across Episodes of Care; AQA Principles of Efficiency Measures).

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

Outcome

☒ Outcome: Click here to name the health outcome

☐ Patient-reported outcome (PRO): Click here to name the PRO

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

☐ Intermediate clinical outcome (*e.g., lab value*): Click here to name the intermediate outcome

☐ Process: Click here to name what is being measured

☐ Appropriate use measure: Click here to name what is being measured

☐ Structure: Click here to name the structure

☐ Composite: Click here to name what is being measured

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

Unplanned hospital visits following ambulatory surgical center (ASC) surgical procedures often reflect procedure-related adverse events and quality issues. Strategies and interventions that have been shown to reduce unplanned hospital visits after outpatient surgical procedures include:

1) Appropriate patient selection for surgical procedures [1];
2) Appropriate patient education on preparation prior to procedures [2];
3) Improving the technical quality of the surgery, including the choice of procedural technique and anesthesia [3];
4) Prevention of surgical site infections through evidence-based guideline-concordant care [4,5]; and
5) Prevention of adverse drug events through medication reconciliation [6].

The measure will identify ASCs that have significantly higher rates of unplanned hospital visits relative to other ASCs performing the same types of surgical procedures on similar patients and will prompt ASCs to evaluate care processes and implement quality improvement strategies.

Citations:

1. Fleisher LA, Pasternak LR, Lyles A. A novel index of elevated risk of inpatient hospital admission immediately following outpatient surgery. *Arch Surg.* 2007;142(3):263-268.
2. Romero A, Joshi GP. Adult Patient for Ambulatory Surgery: Are There Any Limits? *ASA Newsletter.* 2014;78(9):18-20.
3. Whippey A, Kostandoff G, Paul J, Ma J, Thabane L, Ma HK. Predictors of unanticipated admission following ambulatory surgery: a retrospective case-control study. *Can J Anaesth* . 2013;60(7):675-683.
4. Mangram AJ, Horan TC, Pearson ML, Silver LC, Jarvis WR, Committee HICPA. Guideline for prevention of surgical site infection, 1999. *Am J Infect Control.* 1999;27(2):97-134.
5. Agency for Healthcare Research and Quality. Proactive Risk Assessment of Surgical Site Infection in Ambulatory Surgery Centers: Final Contract Report. Chapter 3: Risk-Informed Interventions. April 2013. Available at: http://www.ahrq.gov/research/findings/final-reports/stpra/stpra3.html. Accessed July 18, 2016.
6. Joint Commission. Joint Commission National Patient Safety Goals: Practical Strategies and Helpful Solutions for Meeting these Goals. 2005; http://teacherweb.com/NY/StBarnabas/Law-PublicPolicy/JCINT-2005.pdf. Accessed June 8, 2016.

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

Not applicable. This measure is not derived from patient report.

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

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

The outcome of unplanned hospital visits following outpatient surgery is an accepted measure of outpatient surgical care quality and reflects important features of healthcare structure, process, and service. These features include patient selection and management, technical aspects of the surgery, and delivery of guideline-concordant care. This measure will provide the opportunity for ASCs to become aware of and to lower rates of adverse events leading to hospital visits after general surgery procedures performed at ASCs.

A hospital visit after outpatient surgery is unexpected, and many of the reasons for such hospital visits are preventable. In the literature, hospital visit rates following outpatient surgery vary from 0.5-9.0%, based on the type of surgery, outcome measured (admissions alone or admissions and emergency department [ED] visits), and timeframe for measurement after surgery [1-10]. These hospital visits can occur due to a range of adverse events, including major adverse events, such as infection, post-operative bleeding, and urinary retention. Patients also frequently report minor adverse events – for example, uncontrolled pain, nausea, and vomiting – that may result in unplanned acute care visits following surgery.

There is literature providing evidence that interventions can improve patient outcomes after outpatient surgery. Studies, many focusing on surgeries in the hospital outpatient department setting, point to the importance of post-discharge factors, such as ability to manage pain and availability of a responsible caregiver, in reducing poor outcomes [3, 11-15]. The quality of patient selection, patient preparation, post-operative care and post-discharge planning can affect the rate of adverse events and unplanned hospital visits following outpatient surgery [3, 11-12]. The risk of unplanned hospital visits is influenced by various technical aspects of the surgery, including anesthetic technique [11-13] and length of surgery [12]. Although there is limited evidence in the ASC context, these interventions should be applicable in both settings.

Additionally, there are growing efforts to enable ASC providers to systematically address issues of complications of surgical care and communication between providers of adverse events when they occur [16-18]. For example, the Agency for Healthcare Research and Quality (AHRQ) developed a quality improvement collaborative for the 65 ambulatory surgery facilities in 47 states to reduce healthcare-associated infections and surgical harms in ASCs through 1) the use of a surgical safety checklist curriculum, and 2) improved safety culture through teamwork and communication [18]. ASC providers involved in the collaborative concluded that efforts to increase the availability of meaningful data would be beneficial to the accurate assessment of outcomes in the ASC setting, reduce admissions, and would facilitate ASC's ability to follow patients after discharge.

Citations:

1. Majholm BB. Is day surgery safe? A Danish multicentre study of morbidity after 57,709 day surgery procedures. Acta anaesthesiologica Scandinavica. 2012;56(3):323-331.
2. Whippey A, Kostandoff G, Paul J, Ma J, Thabane L, Ma HK. Predictors of unanticipated admission following ambulatory surgery: a retrospective case-control study. Canadian Journal of Anesthesia/Journal canadien d'anesthésie. 2013;60(7):675-683.
3. Fleisher LA, Pasternak LR, Herbert R, Anderson GF. Inpatient hospital admission and death after outpatient surgery in elderly patients: importance of patient and system characteristics and location of care. Arch Surg. 2004;139(1):67-72.
4. Coley KC, Williams BA, DaPos SV, Chen C, Smith RB. Retrospective evaluation of unanticipated admissions and readmissions after same day surgery and associated costs. Journal of clinical anesthesia. 2002;14(5):349-353.
5. Bain J, Kelly H, Snadden D, Staines H. Day surgery in Scotland: patient satisfaction and outcomes. Quality in Health Care. 1999;8(2):86-91.
6. Fortier J, Chung F, Su J. Unanticipated admission after ambulatory surgery--a prospective study. Canadian journal of anaesthesia = Journal canadien d'anesthesie. 1998;45(7):612-619.
7. Aldwinckle R, Montgomery J. Unplanned admission rates and postdischarge complications in patients over the age of 70 following day case surgery. Anaesthesia. 2004;59(1):57-59.

8. Owens PL, Barrett ML, Raetzman S, Maggard-Gibbons M, Steiner CA. Surgical site infections following ambulatory surgery procedures. *JAMA.* 2014;311 (7): 709-716.
9. Mioton LM, Buck DW, 2nd, Rambachan A, Ver Halen J, Dumanian GA, Kim JY. Predictors of readmission after outpatient plastic surgery. *Plastic & Reconstructive Surgery.* 2014; 133(1):173-180.
10. Bhattacharyya N. Unplanned revisits and readmissions after ambulatory sinonasal surgery. *Laryngoscope.* 2014; 124(9):1983-1987.
11. Romero A, Joshi GP. Adult Patient for Ambulatory Surgery: Are There Any Limits? *ASA Newsletter.* 2014;78(9):18-20.
12. Whippey A, Kostandoff G, Paul J, Ma J, Thabane L, Ma HK. Predictors of unanticipated admission following ambulatory surgery: a retrospective case-control study. *Can J Anaesth.* 2013;60(7):675-683.
13. Perrot DH. Anesthesia Outside the Operating Room in the Office-Based Setting. *Curr Opin Anaesthesiol*. 2008; 21L 480-485
14. Pearson, Alan, Marilyn Richardson, and Michelle Cairns. ""Best practice" in day surgery units: a review of the evidence." *Ambulatory Surgery* 11.1 (2004): 49-54.
15. Allison, Jan, and Michelle George. "Using preoperative assessment and patient instruction to improve patient safety." *AORN journal* 99.3 (2014): 364-375.
16. Agency for Healthcare Research and Quality. Proactive Risk Assessment of Surgical Site Infection in Ambulatory Surgery Centers: Final Contract Report. Chapter 3: Risk-Informed Interventions. April 2013. Available at: http://www.ahrq.gov/research/findings/final-reports/stpra/stpra3.html. Accessed July 18, 2016.
17. Joint Commission. Joint Commission National Patient Safety Goals: Practical Strategies and Helpful Solutions for Meeting these Goals. 2005; http://teacherweb.com/NY/StBarnabas/Law-PublicPolicy/JCINT-2005.pdf. Accessed June 8, 2016.
18. The Health Research & Educational Trust of the American Hospital Association. AHRQ Safety Program for Ambulatory Surgery, May 2017. Available at: https://www.ahrq.gov/sites/default/files/wysiwyg/professionals/quality-patient-safety/hais/tools/ambulatory-surgery/sections/ambulatory-surgery-report.pdf. Accessed June 23, 2017.

Measure Information

This document contains the information submitted by measure developers/stewards, but is organized according to NQF's measure evaluation criteria and process. The item numbers refer to those in the submission form but may be in a slightly different order here. In general, the item numbers also reference the related criteria (e.g., item 1b.1 relates to sub criterion 1b).

Brief Measure Information

NQF #: 3357
Corresponding Measures:
De.2. **Measure Title:** Facility-Level 7-Day Hospital Visits after General Surgery Procedures Performed at Ambulatory Surgical Centers
Co.1.1. **Measure Steward:** Centers for Medicare & Medicaid Services (CMS)
De.3. **Brief Description of Measure:** Facility-level risk-standardized rate of acute, unplanned hospital visits within 7 days of a general surgery procedure performed at an ambulatory surgical center (ASC) among Medicare Fee-For-Service (FFS) patients aged 65 years and older. An unplanned hospital visit is defined as an emergency department (ED) visit, observation stay, or unplanned inpatient admission.
1b.1. **Developer Rationale:** This measure aims to reduce adverse patient outcomes associated with ASC surgeries and improve follow-up care by capturing and illuminating, for providers and patients, post-surgery unplanned hospital visits that are often not visible to providers at ASCs. The measure score will assess quality and inform quality improvement.

S.4. **Numerator Statement:** The outcome being measured is acute, unplanned hospital visits (ED visit, observation stay, or unplanned inpatient admission) occurring within 7 days of a general surgery procedure performed at an ASC.
S.6. **Denominator Statement:** Target Population

Included patients:
The target population for this measure is Medicare FFS patients aged 65 years and older, who are undergoing outpatient general surgery procedures in ASCs that are within the scope of general surgery training. Specifically, the cohort of procedures includes the following types of surgeries: abdominal, alimentary tract, breast, skin/soft tissue, wound, and varicose vein.

The Medicare FFS population was chosen because of the availability of a national dataset (Medicare claims) that could be used to develop, test, and publicly report the measure. We limit the measure to patients who have been enrolled in Medicare FFS Parts A and B for the 12 months prior to the date of surgery to ensure that we have adequate data for identifying comorbidities for risk adjustment.

Included procedures:
The target group of procedures is surgical procedures that (1) are routinely performed at ASCs, (2) involve risk of post-surgery hospital visits, and (3) are within the scope of general surgery training. The scope of general surgery overlaps with that of other specialties (for example, vascular surgery and, plastic surgery). For this measure, we targeted surgeries that general surgeons are trained to perform with the understanding that other subspecialists may also be performing many of these surgeries at ASCs. Since the type of surgeon performing a particular procedure may vary across ASCs in ways that affect quality, the measure is neutral to surgeons' specialty training.

To identify eligible ASC general surgery procedures, we first identified a list of procedures from Medicare's 2014 and 2015 ASC lists of covered procedures, which include procedures for which ASCs can be reimbursed under the ASC payment system. This lists of surgeries is publicly available at: https://www.cms.gov/medicare/medicare-fee-for-service-payment/ascpayment/11_addenda_updates.html (download January 2014 and January 2015 ASC Approved HCPCS Code and Payment Rates, Addendum AA). Surgeries on the ASC list of covered procedures do not involve or require: major or prolonged invasion of body cavities, extensive blood loss, major blood vessels, or care that is either emergent or life-threatening. The ASC list is annually reviewed and updated by Medicare, and includes a transparent public comment submission and review process for addition and/or removal of procedure codes. Using an existing, defined list of surgeries, rather than defining surgeries de novo, is useful for long-term measure maintenance. Procedures listed in Medicare's list of covered ASC procedures are defined using Healthcare Common Procedure Coding System (HCPCS) and Common Procedural Terminology (CPT®) codes.

Ambulatory procedures include a heterogeneous mix of non-surgical procedures, minor surgeries, and more substantive surgeries. The measure is not intended to include very low-risk (minor) surgeries or non-surgical procedures, which typically have a high volume and a very low outcome rate. Therefore, to focus the measure only on the subset of surgeries on Medicare's list of covered ASC procedures that impose a meaningful risk of post-procedure hospital visits, the measure includes only "major" and "minor" procedures, as indicated by the Medicare Physician Fee Schedule global surgery indicator (GSI) values of 090 and 010, respectively. The GSI code reflects the number of post-operative days that are included in a given procedure's global surgical payment and identifies surgical procedures of greater complexity and follow-up care. This list of GSI values is publicly available for calendar year (CY) 2014 at: https://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/PhysicianFeeSched/PFS-Federal-Regulation-Notices-Items/CMS-1600-FC.html and for CY 2015 at: https://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/PhysicianFeeSched/PFS-Federal-Regulation-Notices-Items/CMS-1612-FC.html (download PFS Addenda, Addendum B).

Finally, to identify the subset of general surgery ASC procedures, we reviewed with consultants and Technical Expert Panel (TEP) members the Clinical Classifications Software (CCS) categories of procedures developed by the Agency for Healthcare Research and Quality (AHRQ). We identified and included CCS categories within the scope of general surgery, and only included individual procedures within the CCS categories at the procedure (CPT® code) level if they were within the scope of general surgery practice. We did not include in the measure gastrointestinal endoscopy, endocrine, or vascular procedures, other than varicose vein procedures, because reasons for hospital visits are typically related to patients' underlying comorbidities.

See the attached Data Dictionary, sheet S.9 "Codes Used to Define Cohort" for a complete list of all CPT procedure codes included in the measure cohort.
S.8. **Denominator Exclusions:** The measure excludes surgeries for patients without 7 or more days of continuous enrollment in Medicare FFS Parts A and B after the surgery. The measure excludes these patients to ensure all patients have full data available for outcome assessment.

De.1. **Measure Type:** Outcome
S.17. **Data Source:** Claims, Enrollment Data
S.20. **Level of Analysis:** Facility

IF Endorsement Maintenance – Original Endorsement Date: Most Recent Endorsement Date:

IF this measure is included in a composite, NQF Composite#/title:

IF this measure is paired/grouped, NQF#/title:

De.4. **IF PAIRED/GROUPED, what is the reason this measure must be reported with other measures to appropriately interpret results?** Not applicable

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

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

1a. Evidence to Support the Measure Focus – See attached Evidence Submission Form
Gen_Surg_ASC__NQF_Evidence_Attachment_FINAL_111417.docx
1a.1 For Maintenance of Endorsement: Is there new evidence about the measure since the last update/submission?
Do not remove any existing information. If there have been any changes to evidence, the Committee will consider the new evidence. Please use the most current version of the evidence attachment (v7.1). Please use red font to indicate updated evidence.

1b. Performance Gap

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

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

1b.1. **Briefly explain the rationale for this measure** *(e.g., how the measure will improve the quality of care, the benefits or improvements in quality envisioned by use of this measure)*
If a COMPOSITE (e.g., combination of component measure scores, all-or-none, any-or-none), SKIP this question and answer the composite questions.
This measure aims to reduce adverse patient outcomes associated with ASC surgeries and improve follow-up care by capturing and illuminating, for providers and patients, post-surgery unplanned hospital visits that are often not visible to providers at ASCs. The measure score will assess quality and inform quality improvement.

1b.2. **Provide performance scores on the measure as specified (current and over time) at the specified level of analysis**. *(This is required for maintenance of endorsement. Include mean, std dev, min, max, interquartile range, scores by decile. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities include.) This information also will be used to address the sub-criterion on improvement (4b1) under Usability and Use.*
We assessed ASC-level variation in performance scores using 100% Medicare FFS claims data for 2014-2015 (please see Measure Testing Form Section 1.2 and Section 1.7 for full description of the dataset). Using the 2014-2015 data (which included 286,999 general surgeries from 1,642 ASCs meeting a minimum volume threshold of at least 25 cases), we found variation in the risk-adjusted measure scores among ASCs. The median RSHVR was 0.97, ranging from 0.42 to 2.13 (the 25th and 75th percentiles were 0.90 and 1.10, respectively).

1b.3. **If no or limited performance data on the measure as specified is reported in 1b2, then provide a summary of data from the literature that indicates opportunity for improvement or overall less than optimal performance on the specific focus of measurement.**
Not applicable. We provide performance scores in 1b.2. See Evidence Form for summary of data from the literature that further indicates opportunity for improvement.

1b.4. **Provide disparities data from the measure as specified (current and over time) by population group, e.g., by race/ethnicity, gender, age, insurance status, socioeconomic status, and/or disability.** *(This is required for maintenance of endorsement. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included.) For measures that show high levels of performance, i.e., "topped out", disparities data may demonstrate an opportunity for improvement/gap in care for certain sub-populations. This information also will be used to address the sub-criterion on improvement (4b1) under Usability and Use.*
To examine the impact of social risk factors on the measure calculation we evaluated three indicators of social risk: 1) Medicare-Medicaid dual eligibility (yes vs. no), 2) race (African American vs. all others), and 3) the AHRQ SES Index (explained in Section 2.b.3 of the Measure Testing Form). For these analyses we used 100% Medicare FFS claims data from CYs 2014-2015. These data

included 3,653 ASC facilities and 303,220 general surgery procedures. Our goal for these analyses were twofold: 1) to examine whether these factors were associated with increased risk in hospital visits after adjusting for other risk factors and 2) to evaluate the impact of social risk factors on ASC-level measure scores.

We present these analyses and results in greater detail in Section 2b3.4b of the Measure Testing Form. In brief, to evaluate the association of these risk factors with the outcome, we first quantified the observed rate. We then evaluated the magnitude of association of these social risk factors with the outcome after adjustment for clinical comorbidities, procedure type, and age by including each individual indicator as a variable in our risk-adjustment model. Each factor's effect was quantified using odds ratios (ORs) and tested for significance. In addition, we evaluated the change in each model's predictive ability (c-statistic).

To evaluate the impact of social risk factors on the ASC-level measure scores, we compared RSHVRs calculated with and without each disparity marker included in the model. For these analyses, we calculated the RSHVR difference for each ASC (RSVHR with the social risk variable and RSHVR without the social risk variable) and calculated Pearson correlation coefficients for the paired scores.

We further examined the potential impact of these social risk factors on measure scores by comparing RSHVR distributions using current specifications. ASCs were stratified by the proportion of patients at the ASC with each social risk factor, and placed into quartiles based on these proportions. These stratified distributions were examined for systematic differences in RSHVR across quartiles.

Results
Observed hospital visit rates were higher for patients with each disparity marker: 3.7% for dual-eligible patients compared to 2.2% for non-dual-eligible patients, 3.1% for African-American patients compared to 2.2% for non-African-American patients, and 2.7% for low SES patients (scores below 42.7 on the AHRQ SES Index) compared to 2.2% for higher SES patients (scores above 42.7 on the AHRQ SES index). Furthermore, inclusion of each of these risk factors in our models indicated a statistically significant association after controlling for other risk adjusters in our model (dual-eligible: OR: 1.34, 95% CI: 1.22 -1.48, $p < 0.0001$; race: OR: 1.23, 95% CI: 1.06-1.42, p=0.005; AHRQ SES Index: OR: 1.14, 95% CI: 1.06-1.22, p=0.0004).

However, entering these variables into the risk-adjustment model did not improve model performance (c-statistics remained unchanged) and did not substantially change ASC-level measure scores. Correlation coefficients between risk-standardized hospital visit ratios with and without adjustment for these factors were near 1.

Further, the analyses of ASCs stratified into quartiles based on proportions of dual-eligible, African-American, and low SES patients (as identified by the AHRQ SES Index) showed largely overlapping distributions of the RSHVRs by quartile, although longer tails at the upper ends of the distributions were observed for ASCs with the highest percent of patients with the social risk factor (4th quartile). Distributions for low % of social risk factor ASCs (1st quartile) and high % social risk factor ASCs (4th quartile) by each social risk factor are shown in Table 2, Section 2b3.4b, of the Measure Testing Form.

Based on these analyses we conclude that although the three social risk factors we examined have a modest but statistically significant association with the risk of a hospital visit, these patient-level factors have a limited effect on the ASC-level measure scores. We did not adjust the models for these social risk factors since the association of these factors with the outcome may be quality related, and since these factors have a limited relationship to the facility-level scores.

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

2. Reliability and Validity—Scientific Acceptability of Measure Properties

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

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

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

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

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

S.1. **Measure-specific Web Page** *(Provide a URL link to a web page specific for this measure that contains current detailed specifications including code lists, risk model details, and supplemental materials. Do not enter a URL linking to a home page or to general information.)*
https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/HospitalQualityInits/Measure-Methodology.html

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

S.2b. **Data Dictionary, Code Table, or Value Sets** *(and risk model codes and coefficients when applicable) must be attached. (Excel or csv file in the suggested format preferred - if not, contact staff)*
Attachment **Attachment:** Gen_Surg_ASC_NQF_Data_Dictionary_v1.0.xlsx

S.2c. Is this an instrument-based measure (i.e., data collected via instruments, surveys, tools, questionnaires, scales, etc.)? Attach copy of instrument if available.
No, this is not an instrument-based measure **Attachment:**

S.2d. Is this an instrument-based measure (i.e., data collected via instruments, surveys, tools, questionnaires, scales, etc.)? Attach copy of instrument if available.
Not an instrument-based measure

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

S.3.2. **For maintenance of endorsement,** please briefly describe any important changes to the measure specifications since last measure update and explain the reasons.
Not applicable.

S.4. **Numerator Statement** *(Brief, narrative description of the measure focus or what is being measured about the target population, i.e., cases from the target population with the target process, condition, event, or outcome) DO NOT include the rationale for the measure.*
IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm (S.14).
The outcome being measured is acute, unplanned hospital visits (ED visit, observation stay, or unplanned inpatient admission) occurring within 7 days of a general surgery procedure performed at an ASC.

S.5. **Numerator Details** *(All information required to identify and calculate the cases from the target population with the target process, condition, event, or outcome such as definitions, time period for data collection, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*

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

Outcome Definition

The outcome is unplanned hospital visits, defined as an ED visit, observation stay, or unplanned inpatient admission, occurring within 7 days of the general surgery procedure performed at an ASC identified using Centers for Medicare & Medicaid Services (CMS) Medicare administrative claims data.

Time Period for Data

Numerator time window: 7 days after ASC procedures for unplanned hospital visits.

Denominator time window: General surgery ASC procedures performed during the measurement period.

Identification of Planned Admissions

The measure outcome includes hospital visits within the first 7 days following the procedure, unless that inpatient admission is deemed a "planned" admission. We applied CMS's Planned Readmission Algorithm Version 4.0 to identified planned admissions [1]. Planned admissions are defined as those planned by providers for anticipated medical treatment or procedures that must be provided in the inpatient setting. CMS seeks to count only unplanned admissions in the measure outcome because variation in planned admissions does not reflect quality differences. The algorithm (see the flowchart in the Data Dictionary, first tab, "S.6 Planned Adm Alg Flowchart") identifies inpatient admissions that are typically planned and may occur after the patient's index general surgery procedure, considering a few specific, limited types of care as "planned"(e.g., major organ transplant, rehabilitation, or maintenance chemotherapy). Otherwise, the algorithm defines a planned admission as a non-acute inpatient admission for a scheduled procedure (e.g., total hip replacement or cholecystectomy), and the algorithm never considers inpatient admissions for acute illness or for complications of care planned. The algorithm considers inpatient admissions that include potentially planned procedures with acute diagnoses, or with diagnoses that might represent complications of a surgery, as "unplanned" and thus counts these inpatient admissions in the measure outcome.

Details of the planned admission algorithm and codes to identify planned admissions are in the attached Data Dictionary sheet labeled "S.6 Planned Adm Alg."

Definition of ED Visits and Observation Stay

The measure defines ED visits and observation stays using one of the specified billing codes or revenue center codes identified in Medicare Part B Outpatient hospital claims.

The codes used to define ED visits and observation stays are in the attached Data Dictionary sheet labeled "S.6 Numerator-ED Obs Def."

Citations

1. Horwitz L, Grady J, Cohen D, et al. Development and validation of an algorithm to identify planned readmissions from claims data. Journal of Hospital Medicine. Oct 2015;10(10):670-677.

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

Target Population

Included patients:

The target population for this measure is Medicare FFS patients aged 65 years and older, who are undergoing outpatient general surgery procedures in ASCs that are within the scope of general surgery training. Specifically, the cohort of procedures includes the following types of surgeries: abdominal, alimentary tract, breast, skin/soft tissue, wound, and varicose vein.

The Medicare FFS population was chosen because of the availability of a national dataset (Medicare claims) that could be used to develop, test, and publicly report the measure. We limit the measure to patients who have been enrolled in Medicare FFS Parts A and B for the 12 months prior to the date of surgery to ensure that we have adequate data for identifying comorbidities for risk adjustment.

Included procedures:

The target group of procedures is surgical procedures that (1) are routinely performed at ASCs, (2) involve risk of post-surgery hospital visits, and (3) are within the scope of general surgery training. The scope of general surgery overlaps with that of other specialties (for example, vascular surgery and, plastic surgery). For this measure, we targeted surgeries that general surgeons are trained to perform with the understanding that other subspecialists may also be performing many of these surgeries at ASCs. Since the type of surgeon performing a particular procedure may vary across ASCs in ways that affect quality, the measure is neutral to surgeons' specialty training.

To identify eligible ASC general surgery procedures, we first identified a list of procedures from Medicare's 2014 and 2015 ASC lists of covered procedures, which include procedures for which ASCs can be reimbursed under the ASC payment system. This lists of surgeries is publicly available at: https://www.cms.gov/medicare/medicare-fee-for-service-payment/ascpayment/11_addenda_updates.html (download January 2014 and January 2015 ASC Approved HCPCS Code and Payment Rates, Addendum AA). Surgeries on the ASC list of covered procedures do not involve or require: major or prolonged invasion of body cavities, extensive blood loss, major blood vessels, or care that is either emergent or life-threatening. The ASC list is annually reviewed and updated by Medicare, and includes a transparent public comment submission and review process for addition and/or removal of procedure codes. Using an existing, defined list of surgeries, rather than defining surgeries de novo, is useful for long-term measure maintenance. Procedures listed in Medicare's list of covered ASC procedures are defined using Healthcare Common Procedure Coding System (HCPCS) and Common Procedural Terminology (CPT®) codes.

Ambulatory procedures include a heterogeneous mix of non-surgical procedures, minor surgeries, and more substantive surgeries. The measure is not intended to include very low-risk (minor) surgeries or non-surgical procedures, which typically have a high volume and a very low outcome rate. Therefore, to focus the measure only on the subset of surgeries on Medicare's list of covered ASC procedures that impose a meaningful risk of post-procedure hospital visits, the measure includes only "major" and "minor" procedures, as indicated by the Medicare Physician Fee Schedule global surgery indicator (GSI) values of 090 and 010, respectively. The GSI code reflects the number of post-operative days that are included in a given procedure's global surgical payment and identifies surgical procedures of greater complexity and follow-up care. This list of GSI values is publicly available for calendar year (CY) 2014 at: https://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/PhysicianFeeSched/PFS-Federal-Regulation-Notices-Items/CMS-1600-FC.html and for CY 2015 at: https://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/PhysicianFeeSched/PFS-Federal-Regulation-Notices-Items/CMS-1612-FC.html (download PFS Addenda, Addendum B).

Finally, to identify the subset of general surgery ASC procedures, we reviewed with consultants and Technical Expert Panel (TEP) members the Clinical Classifications Software (CCS) categories of procedures developed by the Agency for Healthcare Research and Quality (AHRQ). We identified and included CCS categories within the scope of general surgery, and only included individual procedures within the CCS categories at the procedure (CPT® code) level if they were within the scope of general surgery practice. We did not include in the measure gastrointestinal endoscopy, endocrine, or vascular procedures, other than varicose vein procedures, because reasons for hospital visits are typically related to patients' underlying comorbidities.

See the attached Data Dictionary, sheet S.9 "Codes Used to Define Cohort" for a complete list of all CPT procedure codes included in the measure cohort.

S.7. Denominator Details *(All information required to identify and calculate the target population/denominator such as definitions, time period for data collection, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b.)*
IF an OUTCOME MEASURE, describe how the target population is identified. Calculation of the risk-adjusted outcome should be described in the calculation algorithm (S.14).
Target Population

Included patients:
The target population for this measure is Medicare FFS patients aged 65 years and older, who are undergoing outpatient general surgery procedures in ASCs that are within the scope of general surgery training. Specifically, the cohort of procedures includes the following types of surgeries: abdominal, alimentary tract, breast, skin/soft tissue, wound, and varicose vein.
The Medicare FFS population was chosen because of the availability of a national dataset (Medicare claims) that could be used to develop, test, and publicly report the measure. We limit the measure to patients who have been enrolled in Medicare FFS Parts A and B for the 12 months prior to the date of surgery to ensure that we have adequate data for identifying comorbidities for risk adjustment.

Included procedures:
The target group of procedures is surgical procedures that (1) are routinely performed at ASCs, (2) involve risk of post-surgery hospital visits, and (3) are within the scope of general surgery training. The scope of general surgery overlaps with that of other specialties (for example, vascular surgery and, plastic surgery). For this measure, we targeted surgeries that general surgeons are trained to perform with the understanding that other subspecialists may also be performing many of these surgeries at ASCs. Since the type of surgeon performing a particular procedure may vary across ASCs in ways that affect quality, the measure is neutral to surgeons' specialty training.

To identify eligible ASC general surgery procedures, we first identified a list of procedures from Medicare's 2014 and 2015 ASC lists of covered procedures, which include procedures for which ASCs can be reimbursed under the ASC payment system. This lists of surgeries is publicly available at: https://www.cms.gov/medicare/medicare-fee-for-service-payment/ascpayment/11_addenda_updates.html (download January 2014 and January 2015 ASC Approved HCPCS Code and Payment Rates, Addendum AA). Surgeries on the ASC list of covered procedures do not involve or require: major or prolonged invasion of body cavities, extensive blood loss, major blood vessels, or care that is either emergent or life-threatening. The ASC list is annually reviewed and updated by Medicare, and includes a transparent public comment submission and review process for addition and/or removal of procedure codes. Using an existing, defined list of surgeries, rather than defining surgeries de novo, is useful for long-term measure maintenance. Procedures listed in Medicare's list of covered ASC procedures are defined using Healthcare Common Procedure Coding System (HCPCS) and Common Procedural Terminology (CPT®) codes.

Ambulatory procedures include a heterogeneous mix of non-surgical procedures, minor surgeries, and more substantive surgeries. The measure is not intended to include very low-risk (minor) surgeries or non-surgical procedures, which typically have a high volume and a very low outcome rate. Therefore, to focus the measure only on the subset of surgeries on Medicare's list of covered ASC procedures that impose a meaningful risk of post-procedure hospital visits, the measure includes only "major" and "minor" procedures, as indicated by the Medicare Physician Fee Schedule global surgery indicator (GSI) values of 090 and 010, respectively. The GSI code reflects the number of post-operative days that are included in a given procedure's global surgical payment and identifies surgical procedures of greater complexity and follow-up care. This list of GSI values is publicly available for calendar year (CY) 2014 at: https://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/PhysicianFeeSched/PFS-Federal-Regulation-Notices-Items/CMS-1600-FC.html and for CY 2015 at: https://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/PhysicianFeeSched/PFS-Federal-Regulation-Notices-Items/CMS-1612-FC.html (download PFS Addenda, Addendum B).
Finally, to identify the subset of general surgery ASC procedures, we reviewed with consultants and Technical Expert Panel (TEP) members the Clinical Classifications Software (CCS) categories of procedures developed by the Agency for Healthcare Research and Quality (AHRQ). We identified and included CCS categories within the scope of general surgery, and only included individual procedures within the CCS categories at the procedure (CPT® code) level if they were within the scope of general surgery practice. We did not include in the measure gastrointestinal endoscopy, endocrine, or vascular procedures, other than varicose vein procedures, because reasons for hospital visits are typically related to patients' underlying comorbidities.

See the attached Data Dictionary, sheet S.9 "Codes Used to Define Cohort" for a complete list of all CPT procedure codes included in the measure cohort.

S.8. **Denominator Exclusions** *(Brief narrative description of exclusions from the target population)*
The measure excludes surgeries for patients without 7 or more days of continuous enrollment in Medicare FFS Parts A and B after the surgery. The measure excludes these patients to ensure all patients have full data available for outcome assessment.

S.9. **Denominator Exclusion Details** *(All information required to identify and calculate exclusions from the denominator such as definitions, time period for data collection, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b.)*
Lack of 7 or more days of continuous enrollment in Medicare FFS after the ASC surgery is determined by patient enrollment status in FFS Parts A and B using the Medicare enrollment file (unless lack of enrollment was due to death). The procedure must be 7 or more days from the end of the month or the enrollment indicators must be appropriately marked for the month that falls within 7 days of the procedure date (unless disenrollment is due to death), otherwise the procedure is excluded.

S.10. **Stratification Information** *(Provide all information required to stratify the measure results, if necessary, including the stratification variables, definitions, specific data collection items/responses, code/value sets, and the risk-model covariates and coefficients for the clinically-adjusted version of the measure when appropriate – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format with at S.2b.)*
Not applicable.

S.11. **Risk Adjustment Type** (Select type. Provide specifications for risk stratification in measure testing attachment)
Statistical risk model
If other:

S.12. **Type of score:**
Ratio
If other:

S.13. **Interpretation of Score** *(Classifies interpretation of score according to whether better quality is associated with a higher score, a lower score, a score falling within a defined interval, or a passing score)*
Better quality = Lower score

S.14. **Calculation Algorithm/Measure Logic** (*Diagram or describe the calculation of the measure score as an ordered sequence of steps including identifying the target population; exclusions; cases meeting the target process, condition, event, or outcome; time period for data, aggregating data; risk adjustment; etc.)*
The measure uses a two-level hierarchical logistic regression model to estimate ASC-level risk-standardized hospital visit ratios (RSHVRs). This approach accounts for the clustering of patients within ASCs and variation in sample size across ASCs. The RSHVR is calculated as the ratio of the predicted to the expected number of post-surgical unplanned hospital visits among ASC's patients. For each ASC, the numerator of the ratio is the number of hospital visits predicted for the ASC's patients, accounting for its observed rate, the number and complexity of general surgery procedures performed at the ASC, and the case mix. The denominator is the number of hospital visits expected nationally for the ASC's case/procedure mix. To calculate an ASC's predicted-to-expected (P/E) ratio, the measure uses a two-level hierarchical logistic regression model. The log-odds of the outcome for an index procedure is modeled as a function of the patient demographic, comorbidity, procedure characteristics, and a random ASC-specific intercept. A ratio greater than one indicates that the ASC's patients have more visits than expected, compared to an average ASC with similar patient and procedural complexity. A ratio less than one indicates that the ASC's patients have fewer post-surgical visits than expected, compared to an average ASC with similar patient and procedural complexity. This approach is analogous to an observed-to-expected ratio, but accounts for within-facility correlation of the observed outcome and sample size differences and accommodates the assumption that underlying differences in quality across ASCs lead to systematic differences in outcomes, and is tailored to and appropriate for a publicly reported outcome measure as articulated in published scientific guidelines [1-3].
Please see Appendix D of the attached technical report for details.

Citations
1. Normand S-LT, Shahian DM. Statistical and clinical aspects of hospital outcomes profiling. Statistical Science. 2007;22(2):206-226.
2. Krumholz HM, Brindis RG, Brush JE, et al. Standards for Statistical Models Used for Public Reporting of Health Outcomes An American Heart Association Scientific Statement From the Quality of Care and Outcomes Research Interdisciplinary Writing Group: Cosponsored by the Council on Epidemiology and Prevention and the Stroke Council Endorsed by the American College of Cardiology Foundation. Circulation. 2006;113(3):456-462.
3. National Quality Forum. Measure Evaluation Criteria and Guidance for Evaluating Measures for Endorsement. 2015; http://www.qualityforum.org/Measuring_Performance/Submitting_Standards/2015_Measure_Evaluation_Criteria.aspx. Accessed July 26, 2016.

S.15. **Sampling** *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*
IF an instrument-based performance measure (e.g., PRO-PM), identify whether (and how) proxy responses are allowed.
Not applicable. This measure is not based on a sample or survey.

S.16. **Survey/Patient-reported data** *(If measure is based on a survey or instrument, provide instructions for data collection and guidance on minimum response rate.)*
Specify calculation of response rates to be reported with performance measure results.
Not applicable. This measure is not based on a sample or survey.

S.17. **Data Source** *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*
If other, please describe in S.18.
Claims, Enrollment Data

S.18. **Data Source or Collection Instrument** *(Identify the specific data source/data collection instrument (e.g. name of database, clinical registry, collection instrument, etc., and describe how data are collected.)*
IF instrument-based, identify the specific instrument(s) and standard methods, modes, and languages of administration.
Medicare administrative claims and enrollment data.

S.19. **Data Source or Collection Instrument** *(available at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
No data collection instrument provided

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

S.21. **Care Setting** *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*
Outpatient Services
If other:

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

2. Validity – See attached Measure Testing Submission Form
Gen_Surg_ASC_NQF_Testing_Attachment_FINAL2_111917.docx

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

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

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

NATIONAL QUALITY FORUM—Measure Testing (subcriteria 2a2, 2b1-2b6)

Measure Number (*if previously endorsed*)**:**
Measure Title: Facility-Level 7-Day Hospital Visits after General Surgery Procedures Performed at Ambulatory Surgical Centers
Date of Submission:

Type of Measure:

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

Instructions

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

<u>Note</u>: The information provided in this form is intended to aid the Standing Committee and other stakeholders in understanding to what degree the testing results for this measure meet NQF's evaluation criteria for testing.

2a2. Reliability testing [10] demonstrates the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise. For **instrument-based measures** (including PRO-PMs) **and composite performance measures**, reliability should be demonstrated for the computed performance score.

2b1. Validity testing [11] demonstrates that the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality. For **instrument-based measures (including PRO-PMs) and composite performance measures**, validity should be demonstrated for the computed performance score.

2b2. Exclusions are supported by the clinical evidence and are of sufficient frequency to warrant inclusion in the specifications of the measure; [12]

AND

If patient preference (e.g., informed decisionmaking) is a basis for exclusion, there must be evidence that the exclusion impacts performance on the measure; in such cases, the measure must be specified so that the information about patient preference and the effect on the measure is transparent (e.g., numerator category computed separately, denominator exclusion category computed separately). [13]

2b3. For outcome measures and other measures when indicated (e.g., resource use):

- **an evidence-based risk-adjustment strategy** (e.g., risk models, risk stratification) is specified; is based on patient factors (including clinical and social risk factors) that influence the measured outcome and are present at start of care; [14,15] and has demonstrated adequate discrimination and calibration

OR

- rationale/data support no risk adjustment/ stratification.

2b4. Data analysis of computed measure scores demonstrates that methods for scoring and analysis of the specified measure allow for **identification of statistically significant and practically/clinically meaningful [16] differences in performance**;

OR

there is evidence of overall less-than-optimal performance.

2b5. If multiple data sources/methods are specified, there is demonstration they produce comparable results.

2b6. Analyses identify the extent and distribution of **missing data** (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias.

Notes

10. Reliability testing applies to both the data elements and computed measure score. Examples of reliability testing for data elements include, but are not limited to: inter-rater/abstractor or intra-rater/abstractor studies; internal consistency for multi-item scales; test-retest for survey items. Reliability testing of the measure score addresses precision of measurement (e.g., signal-to-noise).

11. Validity testing applies to both the data elements and computed measure score. Validity testing of data elements typically analyzes agreement with another authoritative source of the same information. Examples of validity testing of the measure score include, but are not limited to: testing hypotheses that the measures scores indicate quality of care, e.g., measure scores are different for groups known to have differences in quality assessed by another valid quality measure or method; correlation of measure scores with another valid indicator of quality for the specific topic; or relationship to conceptually related measures (e.g., scores on process measures to scores on outcome measures). Face validity of the measure score as a quality indicator may be adequate if accomplished through a systematic and transparent process, by identified experts, and explicitly addresses whether performance scores resulting from the measure as specified can be used to distinguish good from poor quality. The degree of consensus and any areas of disagreement must be provided/discussed.

12. Examples of evidence that an exclusion distorts measure results include, but are not limited to: frequency of occurrence, variability of exclusions across providers, and sensitivity analyses with and without the exclusion.

13. Patient preference is not a clinical exception to eligibility and can be influenced by provider interventions.

14. Risk factors that influence outcomes should not be specified as exclusions.

15. With large enough sample sizes, small differences that are statistically significant may or may not be practically or clinically meaningful. The substantive question may be, for example, whether a statistically significant difference of one percentage point in the percentage of patients who received smoking cessation counseling (e.g., 74 percent v. 75 percent) is clinically meaningful; or whether a statistically significant difference of $25 in cost for an episode of care (e.g., $5,000 v. $5,025) is practically meaningful. Measures with overall less-than-optimal performance may not demonstrate much variability across providers.

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

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

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

Measure Specified to Use Data From: **(*must be consistent with data sources entered in S.17*)**	**Measure Tested with Data From:**

☐ abstracted from paper record	☐ abstracted from paper record
☒ claims	☒ claims
☐ registry	☐ registry
☐ abstracted from electronic health record	☐ abstracted from electronic health record
☐ eMeasure (HQMF) implemented in EHRs	☐ eMeasure (HQMF) implemented in EHRs
☒ other: Enrollment database and denominator files	☒ other: Enrollment database and denominator files

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

The measure requires a data source that allows us to link patient data across care settings to identify appropriate surgical procedures for inclusion, comorbidities for risk adjustment, and the outcome of hospital visits [1-3]. Therefore, we used claims data, as they support these linkages and were available for the population of interest.

1. To develop and test the patient-level model, we used a national dataset of Calendar Year (CY) 2015 Medicare claims data from Health Account Joint Information (HAJI) database that included Medicare Inpatient, Outpatient, and Carrier (Part B Physician) claims.

a. Datasets used to <u>define the cohort</u>:

-Outpatient general surgery procedures performed at Ambulatory Surgical Centers (ASCs) were identified using the full set of Medicare beneficiaries' claims from the CY 2015 Carrier non-institutional claims, which included the ASC facility claim (with a unique facility identifier).
-Enrollment database and denominator files: These datasets contain Medicare Fee-For-Service (FFS) enrollment, demographic, and death information for Medicare beneficiaries, which is used to determine inclusion criteria.

b. Datasets used to <u>capture the outcome</u> (hospital visits):

-The outcomes of emergency department (ED) visits and observation stays after general surgery ASC procedures were identified from the CY 2015 hospital outpatient institutional claims and inpatient hospital admissions from the CY 2015 inpatient institutional claims.

c. Datasets used to <u>identify comorbidities for risk adjustment</u>:

-Inpatient and outpatient claims (institutional and non-institutional carrier) data from the year prior (CY 2014) were used to identify comorbidities for risk adjustment for these patients.

2. To align with the Center for Medicare & Medicaid Services' (CMS's) intention to use more than 1 year of data for public reporting to ensure reliable estimates, we calculated ASCs' measure scores and the measure score reliability for a 2-year reporting period. Specifically, we used 2 years of claims data, which included Medicare Inpatient, Outpatient, and Carrier (Part B Physician) claims for CY 2014 and CY 2015 from the HAJI database, to calculate ASCs' measure scores.

3. We used the American Community Survey data from the United States (US) Census Bureau (years 2009-2013) to derive the Agency for Healthcare Research and Quality (AHRQ) socioeconomic status (SES) index for each zip code in the US. Other social risk factors were identified using enrollment and denominator files described above.

4. To calculate measure score reliability for a 2-year reporting period, we used a 4-year cohort of Medicare claims data from the HAJI database for CYs 2012-2015 (January 1, 2012 – December 31, 2015). We created two patient samples per facility that were equivalent in size to 2 years of data.

The datasets used for testing vary by testing type; see Section 1.7 for details.

References

1. Hosmer DW, Lemeshow S. Introduction to the logistic regression model. *Applied Logistic Regression, Second Edition.* 2000:1-30.
2. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics.* 1988:837-845.
3. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics.* 1977;33(1):159-174.

1.3. What are the dates of the data used in testing?

We used Medicare FFS data from CYs 2011-2015. Years of data vary by testing type.

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

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

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

The number of measured entities varied by testing type; see Section 1.7 for details.

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

The number of patients varied by testing type; see Section 1.7 for details.

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

As described in Section 1.2, we used CY 2015 Medicare claims data from the HAJI database that included Medicare Inpatient, Outpatient, and Carrier (Part B Physician) claims to develop the patient-level model, and CYs 2014-2015 to perform facility-level testing. The measure cohort inclusion and exclusion criteria are specified in the Measure Submission Form, Sections S.7 to S.9.

The datasets, number of measured entities, number of general surgery procedures, and demographic profile for the patients used in each type of testing are as follows:

1. Medicare FFS CY 2015 Dataset

-Dates: January 1, 2015 – December 31, 2015

-Number of facilities: 3,251 ASCs
-Number of general surgery procedures: 149,468
-Demographic characteristics: average age of 76.3 years; 45.73% female
-Dataset used for: defining the cohort, testing the exclusion criteria (Section 2b2.2), disparities testing (Section 2b3.4b)

2. Development Sample and Validation Sample

The 2015 Development and Validation Samples were derived by selecting two random samples from the Medicare FFS CY 2015 Dataset. The Development Sample included 50% of the general surgery ASC procedures in the Medicare FFS CY 2015 Dataset, and the Validation Sample included 50% of the general surgery ASC procedures in the Medicare FFS CY 2015 Dataset.

Development Sample

-Dates: January 1, 2015 – December 31, 2015

-Number of facilities: 2,966 ASCs
-Number of general surgery procedures: 74,734
-Demographic characteristics: average age of 76.3 years; 45.83% female
-Dataset used for: testing data element reliability (Section 2a2.3), testing the patient-level risk-adjustment model (Section 2b3.4a)

Validation Sample

-Dates: January 1, 2015 – December 31, 2015

-Number of facilities: 2,961 ASCs
-Number of general surgery procedures: 74,734
-Demographic characteristics: average age of 76.3 years; 45.62% female
-Dataset used for: testing data element reliability (Section 2a2.3), validating the patient-level risk adjustment model (Section 2a2.3), internal validation of the model (see Section 2b1.3)

3. Medicare FFS CYs 2014-2015 Dataset

-Dates: January 1, 2014 – December 31, 2015

-Number of facilities (with at least 25 cases): 1,642 ASCs

-Number of general surgery procedures (across ASCs with at least 25 cases): 286,999

-Demographic characteristics: average age of 76.4 years; 45.57% female

-Dataset used for: testing facility-level score distribution

4. Medicare FFS CYs 2012-2015 Dataset

-Dates: January 1, 2012 – December 31, 2015

-Number of facilities: 4,177 ASCs

-Number of general surgery procedures: 619,499

-Demographic characteristics: average age of 76.5 years; 46.18 % female

-Dataset used for: testing facility-level reliability

Note: For all cohorts defined above, we use 1 additional year of data (the year prior to the first year) to gather risk-adjustment variables for the patients undergoing procedures in the first year of the cohort (example: for dataset #4, we use calendar year 2011 data to gather risk factors for patients undergoing procedures in 2012).

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

As detailed below and in Section 2b3.4b, we considered two patient-level social risk factor variables (Medicaid dual-eligibility status and African-American race) and a composite measure of low SES (the AHRQ-SES index score). In addition, we examined the facility-level proportions of dual-eligible patients, of African-American patients, and of low-SES patients. These analyses were performed with the Medicare FFS CYs 2014-2015 Dataset and data from the Census Bureau's American Community Survey.

We selected social risk factors to analyze after reviewing the literature and examining available national data sources. In the ambulatory surgery setting, studies have demonstrated higher risk of post-procedure hospital visits for African-American and Hispanic patients and for patients residing in lower-income households [1-4].

Potential pathways for SES and race variables' effects are described below in Section 2b3.3a.

The SES and race variables that we examined are:

- Dual-eligible status
- African-American race
- AHRQ-validated SES index score (summarizing the information from the following variables: percentage of people in the labor force who are unemployed, percentage of people living below poverty level, median household income, median value of owner-occupied dwellings, percentage of people ≥25 years of age with less than a 12th-grade education, percentage of people ≥25 years of age completing ≥4 years of college, and percentage of households that average ≥1 people per room)

In selecting variables, our intent was to be responsive to the National Quality Forum (NQF) guidelines for measure developers and the findings of recent work funded by the Improving Medicare Post-Acute Care Transformation (IMPACT) Act of 2014 [3, 4]. Our approach was to examine patient-level indicators of both SES and race that are reliably available for all Medicare beneficiaries and linkable to claims data and to select those that have established validity.

Previous studies examining the validity of data on patients' race collected by CMS have shown that only the data identifying African-American beneficiaries have adequate sensitivity and specificity to be applied broadly in research or measures of quality. While this variable is not ideal because it groups all non-African-American beneficiaries together, it is currently the only race variable available on all beneficiaries across the nation that is linkable to claims data.

Similarly, we recognize that Medicare-Medicaid dual eligibility has limitations as a proxy for patients' income or assets because it does not provide a range of results and is only a dichotomous measure. However, the eligibility threshold for over 65-year-old Medicare patients is valuable, as it considers both income and assets and is consistently applied across states. For both our race and dual-eligible variables, there is a body of literature demonstrating differential health care and health outcomes among beneficiaries indicating that these variables, while not ideal, allow us to examine some of the pathways of interest [3].

Finally, we selected the AHRQ-validated SES Index score because it is a well-validated variable that describes the average SES of people living in defined geographic areas [5]. Its value as a proxy for patient-level information is dependent on having the most granular-level data with respect to communities in which patients live. We used data from the American Community Survey to create AHRQ SES Index scores at the census block group level and then mapped them to 9-digit ZIP codes via vendor software. The patient-level Medicare FFS claims data were then linked to the AHRQ SES Index scores by patients' ZIP codes. Given the variation in cost of living across the country, we adjusted the median income and median property value components of the AHRQ SES Index by regional price parity values published by the Bureau of Economic Analysis. This provided a better marker of low-SES neighborhoods in high-expense geographic areas.

Citations

1. Bhattacharyya N. Healthcare disparities in revisits for complications after adult tonsillectomy. *Am J Otolaryngol*. 2015 Mar-Apr;36(2):249-253.
2. Menachemi N, Chukmaitov A, Brown LS, et al. Quality of care differs by patient characteristics: outcome disparities after ambulatory surgical procedures. *Am J Med Qual*. 2007 Nov-Dec;22(6):395-401.
3. Department of Health and Human Services, Office of the Assistant Secretary of Planning and Evaluation. Report to Congress: Social Risk factors and Performance Under Medicare's Value-based Payment Programs. 2016; https://aspe.hhs.gov/pdf-report/report-congress-social-risk-factors-and-performance-under-medicares-value-based-purchasing-programs. Accessed November 10, 2017.
4. National Academies of Sciences, Engineering, and Medicine (NASEM); *Accounting for Social Risk Factors in Medicare Payment: Data*. Washington DC: National Academies Press; 2016.
5. Bonito A, Bann C, Eicheldinger C, et al. Creation of new race-ethnicity codes and socioeconomic status (SES) indicators for Medicare beneficiaries. Final report, sub-task. 2008;2.

2a2. RELIABILITY TESTING

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

2a2.1. What level of reliability testing was conducted? (*may be one or both levels*)
☒ **Critical data elements used in the measure** (*e.g., inter-abstractor reliability; data element reliability must address ALL critical data elements*)
☒ **Performance measure score** (e.g., *signal-to-noise analysis*)

2a2.2. For each level checked above, describe the method of reliability testing and what it tests (*describe the steps—do not just name a method; what type of error does it test; what statistical analysis was used*)

Data Element Reliability

In constructing the measure in Medicare FFS patients, we aim to utilize only those data elements from claims data that have both face validity and reliability. We avoid the use of fields that are thought to be coded inconsistently across ASCs. Specifically, we used fields that are consequential for payment and which are audited. We identify such variables through empiric analyses and our understanding of CMS auditing and billing policies, and we seek to avoid variables which do not meet this standard.

In addition, CMS has in place several auditing programs used to assess overall claims coding accuracy, to ensure appropriate billing, and for overpayment recoupment. CMS routinely conducts data analysis to identify potential problem areas and detect fraud and audits important data fields used in our measures, including diagnosis and procedure codes and other elements that are consequential for payment.

Measure Score Reliability

We tested the reliability of the facility measure score by calculating the intra-class correlation coefficient (ICC) of the measure score. To calculate the ICC, we used the Medicare FFS CYs 2012-2015 Dataset. For ASCs with two or more general surgery procedures, these procedures were randomly split into the two samples within each facility. The ASCs with one procedure were randomly split into the two samples. The ICC evaluated the agreement between the risk-standardized hospital visit ratios (RSHVRs) calculated in the two randomly selected samples [1].

Citations

1. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics.* 1977;33(1):159-174.

2a2.3. For each level of testing checked above, what were the statistical results from reliability testing? (e.*g., percent agreement and kappa for the critical data elements; distribution of reliability statistics from a signal-to-noise analysis*)

Data Element Reliability

Table 1: Risk Variable Frequencies, Development and Validation Samples (Medicare 100% FFS Cohort)

Variable (definition)	Development Sample (50%)		Validation Sample (50%)	
	#	%	#	%
N	-	-	-	-
Age: mean (standard deviation [SD])	76.3	7.2	76.3	7.2
Procedure type: Abdomen and its contents	9,506	12.7%	9,474	12.7%
Procedure type: Alimentary tract	4,941	6.6%	5,143	6.9%
Procedure type: Breast	5,089	6.8%	5,094	6.8%
Procedure type: Skin/soft tissue	41,334	55.3%	41,357	55.3%
Procedure type: Wound	13,277	17.8%	13,087	17.5%
Procedure type: Vascular	587	0.8%	579	0.8%
Work Relative Value Units: mean (SD)	7	3.9	7	3.9
Comorbidities	-	-	-	-
Other benign tumors (CC 15, 16)	59,878	80.1%	59,906	80.2%
Liver or biliary disease (CC 27, 28, 29, 30, 31, 32)	6,621	8.9%	6,650	8.9%
Intestinal obstruction or perforation (CC 33)	1,482	2.0%	1,446	1.9%
Dementia or senility (CC 51, 52, 53)	5,611	7.5%	5,697	7.6%
Psychiatric disorders (CC 57, 58, 59, 60, 61, 62, 63)	15,913	21.3%	15,877	21.2%
Other significant central nervous system (CNS) disease (CC 77, 78, 79, 80)	2,698	3.6%	2,745	3.7%

Variable (definition)	Development Sample (50%)		Validation Sample (50%)	
	#	%	#	%
Ischemic heart disease (CC 86, 87, 88, 89)	21,613	28.9%	21,373	28.6%
Specified arrhythmias and other heart rhythm disorders (CC 96, 97)	21,055	28.2%	21,047	28.2%
Stroke (CC 99, 100)	3,215	4.3%	3,273	4.4%
Chronic lung disease (CC 110, 111, 112, 113)	15,192	20.3%	14,976	20.0%
Pneumonia (CC 114, 115, 116)	4,910	6.6%	4,816	6.4%
Dialysis or sever chronic kidney disease (CC 134, 136, 137)	2,122	2.8%	1,990	2.7%
Benign prostatic hyperplasia (ICD-9 codes: 60000, 60001, 60020, 60021, 60090, 6091; ICD-10 codes: N40.0, N40.1, N40.2, N40.3)	14,499	19.4%	14,846	19.9%
Cellulitis, local skin infection (CC 164)	10,371	13.9%	10,541	14.1%
Major traumatic fracture or internal injury (CC 169, 170, 171, 172, 173, 174)	25,337	33.9%	25,389	34.0%
Complications of care (CC 176, 177)	6,083	8.1%	6,179	8.3%
Chronic anticoagulant use (ICD-9 code: V5861; ICD-10 code: Z7901 [long-term [19] use of anticoagulants])	7,671	10.3%	7,653	10.2%
Opioid abuse (ICD-9 codes: 30400, 30401, 30402, 30403, 30470, 30471, 30472, 30403, 30550, 30551, 30552, 30553; ICD-10: codes: F11.10, F11.120, F11.121, F11.122, F11.129, F11.14, F11.150, F11.151, F11.159, F11.181, F11.182, F11.188, F11.19, F11.20, F11.21, F11.220, F11.221, F11.222, F11.229, F11.23, F11.24, F11.250, F11.251, F11.259, F11.281, F11.282, F11.288, F11.29)	386	0.5%	345	0.5%

Measure Score Reliability
Testing measure score reliability yielded an ICC [2,1] of 0.530.

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

Data Element Reliability Results
Table 1 above shows the frequencies across the two split samples for all variables included in the final model. As the results in Table 1 show, the frequencies of the risk variables were similar in the Development and Validation Samples, indicating good variable consistency and data element reliability.

Measure Score Reliability Results
The ICC [2,1] score of 0.530, calculated for four years of data, indicates moderate measure score reliability.

2b1. VALIDITY TESTING

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

☒ **Performance measure score**

☐ **Empirical validity testing**

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

2b1.2. For each level of testing checked above, describe the method of validity testing and what it tests (*describe the steps—do not just name a method; what was tested, e.g., accuracy of data elements compared to authoritative source, relationship to another measure as expected; what statistical analysis was used)*

We demonstrated measure validity through relevant prior validity testing we conducted for other claims-based measures, through the application of established measure development guidelines, and through assessment by external groups.

Validity of Claims-Based Measures
For several other NQF-endorsed measures, our team has demonstrated the validity of using claims data for risk adjustment in lieu of medical record data in estimating facility-level measure scores. CMS has validated six NQF-endorsed measures currently in public reporting (acute myocardial infarction [AMI], heart failure, and pneumonia mortality and readmission measures) with models that used medical record-abstracted data for risk adjustment. Specifically, we conducted claims model validation by building comparable models using abstracted medical record data for risk adjustment for AMI patients (Cooperative Cardiovascular Project data), heart failure patients (National Heart Failure data), and pneumonia patients (National Pneumonia Project dataset). When both models were applied to the same patient population, the hospital risk-standardized rates estimated using the claims-based risk-adjustment models had a high level of agreement with the results based on the medical record model, thus supporting the use of the claims-based models for public reporting. Our group has reported these findings in the peer-reviewed literature [1-6]. While the applicability of these findings to our measure may be limited because these medical record validations were focused on patients admitted for specific medical conditions, they nevertheless suggest claims data generally have an acceptable degree of agreement with clinical data at a facility level.

Validity Indicated by Established Measure Development Guidelines:
We developed this measure in consultation with national guidelines for publicly reported outcome measures, with input from outside experts and the public. The measure is consistent with the technical approach to outcomes measurement set forth in NQF guidance for outcome measures [7], CMS Measure Management System (MMS) guidance, and guidance articulated in the American Heart Association scientific statement entitled, “Standards for Statistical Models Used for Public Reporting of Health Outcomes” [8].

Validity as Assessed by External Groups:
Throughout the measure development process, we obtained expert and stakeholder input through holding regular discussions with external clinical consultants, consulting our national Technical Expert Panel (TEP), and holding a 20-day public comment period.

Yale New Haven Health Services Corporation – Center for Outcomes Research and Evaluation (CORE) clinicians, as well as clinical experts in the field of surgery, met regularly to discuss all aspects of measure development, including the cohort, outcome definition, and risk adjustment.

In addition to the consultations and in alignment with CMS MMS guidance, we convened a TEP to provide input and feedback during measure development from a group of recognized experts in relevant fields. To convene the TEP, we released a public call for nominations and selected individuals to represent a range of perspectives, including clinicians, patients, and individuals with expertise in quality improvement and performance measurement. We held two structured TEP conference calls consisting of presentation of key issues, our proposed approach, and relevant data, followed by open discussion among TEP members. We made modifications to the measure specifications (e.g., cohort definition, risk adjustment) based on TEP feedback on the measure.

Additionally, we held a three-week public comment period to solicit input on the measure's methodology and preliminary specifications. We revised the measure in response to public comment and posted a summary of the comments received as well as the updates made to the measure (available in the Downloads section at https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/MMS/PC-Updates-on-Previous-Comment-Periods.html). This NQF application includes the measure's final specifications, inclusive of the revisions after consideration of the public comments.

Face Validity as Determined by the TEP:
We systematically assessed the face validity of the measure score as an indicator of quality by confidentially soliciting the TEP members' agreement with the following two statements via an online survey following the final TEP meeting:

1. Please rate the following statement on a scale of 1 (strongly agree) to 6 (strongly disagree): "The risk-standardized hospital visit rates obtained from the 'Hospital Visits after General Surgery Ambulatory Surgical Center Procedures' measure as specified are valid and useful measures of ASC general surgical quality of care."

2. Please rate the following statement on a scale of 1 (strongly agree) to 6 (strongly disagree): "The risk-standardized hospital visit rates obtained from the 'Hospital Visits after General Surgery Ambulatory Surgical Center Procedures' measure as specified will provide ASCs with information that can be used to improve their quality of care."

List of TEP Members
1) Robin Blomberg, BA, MA – National Forum of End-Stage Renal Disease, Network 16 (Representative for Kidney Patient Advisory Council); Seattle, WA
2) Kirk Campbell, MD – New York University Hospital for Joint Diseases (Clinical Assistant Professor of Orthopedic Surgery); New York, NY
3) Gary Culbertson, MD, FACS – Iris Surgery Center (Surgeon; Medical Director); Sumter, SC
4) Martha Deed, PhD – Consumers Union Safe Patient Project (Patient Safety Advocate); Austin, TX
5) James Dupree, MD, MPH – University of Michigan (Urologist; Health Services Researcher); Ann Arbor, MI
6) Nester Esnaola, MD, MPH, MBA – Fox Chase Cancer Center (Professor of Surgery; Associate Director for Cancer Health Disparities and Community Engagement); Philadelphia, PA
7) John Gore, MD, MS – University of Washington (Associate Professor of Urology); Seattle, WA
8) Lisa Ishii, MD, MHS – Johns Hopkins School of Medicine (Associate Professor); American Academy of Otolaryngology-Head and Neck Surgery (Coordinator for Research and Quality); Baltimore, MD; Alexandria, VA
9) Atul Kamath, MD – Perelman School of Medicine, University of Pennsylvania (Assistant Professor and Clinical Educator Director of Orthopedic Surgery); Hospital of the University of Pennsylvania (Attending Surgeon); Philadelphia, PA
10) Tricia Meyer, PharmD, MS, FASHP – Scott & White Medical Center (Regional Director of Pharmacy); Texas A&M University College of Medicine (Associate Professor of Anesthesiology); Temple, TX
11) Linda Radach, BA – Consumers Union Safe Patient Project (Patient Safety Advocate); Austin, TX

12) Amita Rastogi, MD, MHA, CHE, MS – Health Care Incentives Improvement Institute (Chief Medical Officer); Newtown, CT
13) Donna Slosburg, RN, BSN, LHRM, CASC – ASC Quality Collaboration (Executive Director); St. Pete Beach, FL
14) Julie Thacker, MD, FACS – Duke Health and Hospital System (Medical Director of Evidence-Based Perioperative Care); Duke School of Medicine Clinical Research Unit (Medical Director, Department of Surgery); Durham, NC
15) Thomas Tsai, MD, MPH – Brigham and Women's Hospital (General Surgeon); Harvard School of Public Health (Research Associate); Boston, MA

Process Used to Identify International Classification of Diseases, Tenth Revision (ICD-10) Codes

This application includes ICD-10 codes that correspond to all International Classification of Diseases, Ninth Revision (ICD-9) codes included in the specifications. The goal was to convert this measure into a new code set, fully consistent with the intent of the original measure.

ICD-10 diagnosis and procedure codes used to define the Planned Admission Algorithm were identified from the 2015 version of the AHRQ Clinical Classification Software (CCS) categories specified for ICD-10, followed by clinician review. The algorithm also includes some individual ICD-9 codes. To create the crosswalk for the ICD-9-level codes, we used the 2015 ICD-9-CM to ICD-10-CM General Equivalence Mappings tool, made available by CMS, followed by team review.

Citations

1. Krumholz HM, Wang Y, Mattera JA, et al. An administrative claims model suitable for profiling hospital performance based on 30-day mortality rates among patients with an acute myocardial infarction. *Circulation*. 2006 Apr 4;113(13):1683-92.
2. Krumholz HM, Lin Z, Drye EE, et al. An administrative claims measure suitable for profiling hospital performance based on 30-day all-cause readmission rates among patients with acute myocardial infarction. *Circ Cardiovasc Qual Outcomes*. 2011 Mar 1;4(2):243-52.
3. Krumholz HM, Wang Y, Mattera JA, et al. An administrative claims model suitable for profiling hospital performance based on 30-day mortality rates among patients with heart failure. *Circulation*. 2006 Apr 4;113(13):1693-701.
4. Keenan PS, Normand S-LT, Lin Z, et al. An administrative claims measure suitable for profiling hospital performance on the basis of 30-day all-cause readmission rates among patients with heart failure. *Circ Cardiovasc Qual Outcomes*. 2008 Sep;1(1):29-37.
5. Bratzler DW, Normand S-LT, Wang Y, et al. An administrative claims model for profiling hospital 30-day mortality rates for pneumonia patients. *PLoS One*. 2011 Apr 12;6(4):e17401.
6. Lindenauer PK, Normand S-LT, Drye EE, et al. Development, validation, and results of a measure of 30-day readmission following hospitalization for pneumonia. *J Hosp Med*. 2011 Mar;6(3):142-50.
7. National Quality Forum. National voluntary consensus standards for patient outcomes, first report for phases 1 and 2: A consensus report http://www.qualityforum.org/projects/Patient_Outcome_Measures_Phases1-2.aspx. Accessed August 19, 2010.
8. Krumholz HM, Brindis RG, Brush JE, et al. Standards for statistical models used for public reporting of health outcomes: An American Heart Association scientific statement from the Quality of Care and Outcomes Research Interdisciplinary Writing Group: Cosponsored by the Council on Epidemiology and Prevention and the Stroke Council endorsed by the American College of Cardiology Foundation. *Circulation.* 2006;113(3):456-462.

2b1.3. What were the statistical results from validity testing? (*e.g., correlation; t-test*)

Face Validity as Determined by the TEP:

14 out of the 15 TEP members responded to the face validity survey. Of the 14 respondents, 12 respondents indicated that they somewhat, moderately, or strongly agreed; and two respondents moderately disagreed with the following two statements:

1. "The risk-standardized hospital visit rates obtained from the Hospital Visits after General Surgery Ambulatory Surgical Center Procedures ASC measure, as specified, are valid and useful measures of ASC general surgical quality of care."

2. "The risk-standardized hospital visit rates obtained from the Hospital Visits after General Surgery Ambulatory Surgical Center Procedures' measure, as specified, will provide ASCs with information that can be used to improve their quality of care."

2b1.4. What is your interpretation of the results in terms of demonstrating validity? (i.*e., what do the results mean and what are the norms for the test conducted?*)

These validity testing results demonstrate TEP agreement with the overall face validity of the measure.

2b2. EXCLUSIONS ANALYSIS

NA ☐ no exclusions — *skip to section 2b3*

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

We determined the single exclusion criterion to be appropriate based on clinical considerations. We examined the overall frequency and proportion of the total cohort excluded for the single exclusion criterion.

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

Applying our inclusion criteria (general surgery procedures, including abdomen and its content, alimentary tract, breast, skin/soft tissue, wound, and varicose vein procedures performed on patients aged ≥65 enrolled in Medicare FFS Parts A and B in the 12 months prior to the date of surgery) to the Medicare FFS CY 2015 Dataset resulted in an initial cohort of 149,512 ASC general surgery procedures. We then applied the following exclusion criterion (see the Measure Submission Form, Sections S.8 and S.9, for exclusion rationale): Excluded surgeries for patients who survived at least 7 days, but were not continuously enrolled in Medicare FFS Parts A and B within 7 days of the general surgery ASC procedure.

This resulted in excluding 44 (0.03%) general surgery procedures performed at ASCs. Thus, the final Medicare FFS CY 2015 Dataset included 149,468 general surgery ASC procedures performed at 3,251 ASCs. Given the few cases affected, we did not examine the distribution of cases across ASCs or the effect of the exclusion on the measure scores.

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

We exclude surgeries for patients without continuous enrollment in Medicare FFS Parts A and B within 7 days of the general surgery ASC procedure. This exclusion is narrowly targeted and necessary to ensure all patients have full data available for outcome assessment. This exclusion criterion removes a small number (0.03%) of general surgery ASC procedures.

2b3. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES
If not an intermediate or health outcome, or PRO-PM, or resource use measure, skip to section 2b4.

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

☐ **No risk adjustment or stratification**

☒ **Statistical risk model with 21 risk factors**

☐ **Stratification by risk categories**

☐ **Other,**

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

The measure uses a two-level hierarchical logistic regression model to estimate ASC-level RSHVRs. This approach accounts for the clustering of patients within ASCs and variation in sample size across ASCs.

The stepwise selection procedure identified age, 17 comorbidities, work Relative Value Units (RVUs) to adjust for surgical procedural complexity, procedure type (abdomen vs. alimentary tract vs. breast vs. skin/soft tissue vs. wound vs. varicose vein), and one interaction term. For the final model, we retained these variables and one variable (opioid use) that had a p-value of 0.0917 because experts advised it was an important risk predictor and expressed a strong preference for including it in the model. Work RVUs are assigned to each Current Procedural Terminology (CPT®) procedure code and approximate surgical procedural complexity by incorporating elements of physician time and effort. For patients with multiple concurrent CPT procedure codes, we risk adjust for the CPT code with the highest Work RVU value.

Model Variables:
1. Age
2. Procedure Type: Abdomen and its contents
3. Procedure Type: Alimentary tract
4. Procedure Type: Breast
5. Procedure Type: Skin/soft tissue
6. Procedure Type: Wound
7. Procedure Type: Varicose vein
8. Work Relative Value Units
9. Other benign tumors (CC 15, 16)
10. Liver or biliary disease (CC 27, 28, 29, 30, 31)
11. Intestinal obstruction or perforation (CC 33)
12. Dementia or senility (CC 51, 52, 53)
13. Psychiatric disorders (CC 57, 58, 59, 60, 61, 62, 63)
14. Other significant central nervous system (CNS) disease (CC 77, 78, 79, 80)
15. Ischemic heart disease (CC 86, 87, 88, 89)

16. Specified arrhythmias and other heart rhythm disorders (CC 96, 97)
17. Stroke (CC 99, 100)
18. Chronic lung disease (CC 110, 111, 112, 113)
19. Pneumonia (CC 114, 115, 116)
20. Dialysis or sever chronic kidney disease (CC 134, 136, 137)
21. Benign prostatic hyperplasia (ICD-9 codes: 60000, 60001, 60020, 60021, 60090, 6091; ICD-10 codes: N40.0, N40.1, N40.2, N40.3)
22. Cellulitis, local skin infection (CC 164)
23. Major traumatic fracture or internal injury (CC 169, 170, 171, 172, 173, 174)
24. Complications of care (CC 176, 177)
25. Chronic anticoagulant use (ICD-9 code: V5861; ICD-10 code: Z7901 [long-term {current} use of anticoagulants])
26. Opioid abuse (ICD-9 codes: 30400, 30401, 30402, 30403, 30470, 30471, 30472, 30403, 30550, 30551, 30552, 30553; ICD-10: codes: F11.10, F11.120, F11.121, F11.122, F11.129, F11.14, F11.150, F11.151, F11.159, F11.181, F11.182, F11.188, F11.19, F11.20, F11.21, F11.220, F11.221, F11.222, F11.229, F11.23, F11.24, F11.250, F11.251, F11.259, F11.281, F11.282, F11.288, F11.29)27
27. Procedure type*RVU

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

Not applicable. This measure is risk adjusted.

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

Our approach to risk adjustment is tailored to, and appropriate for, a publicly reported outcome measure as articulated in published scientific guidelines [1,2]. For example, we only adjust for risk factors that are present at the start of care. We do not risk adjust for conditions that are possible adverse events of care and that are only recorded at the time of the surgery (see Data Dictionary, Sheet 2b4.3 Risk Model Specs). We do not adjust for factors related to the delivery of care that may reflect care quality.

The measure employs a hierarchical logistic regression model (a form of hierarchical generalized linear model [HGLM]) to create an ASC-level 7-day RSHVR. This approach to modeling appropriately accounts for the structure of the data (patients clustered within facilities), the underlying risk due to patients' procedures/comorbidities, and sample size at a given ASC when estimating hospital visit ratios. In brief, the approach simultaneously models two levels (patient and facility) to account for the variance in patient outcomes within and between facilities [2]. At the patient level, the model adjusts the log-odds of hospital visits within 7 days after the procedure for selected demographic, clinical, and procedure risk variables. The second level models the facility-specific intercepts as arising from a normal distribution. The facility intercept, or facility-specific effect, represents the ASC contribution to the risk of 7-day hospital visits, after accounting for patient risk and sample size, and can be inferred as a measure of quality. If there were no differences among ASCs, then after adjusting for patient risk, the facility intercepts would be identical across all ASCs.

<u>Candidate Risk-Adjustment Variables:</u>

The measure adjusts for differences in patient comorbidities, demographics, and in procedure-related differences in risk across ASCs. We identified potential candidate risk factors through: 1) prior work on related quality measures (including the related urology and orthopedic ASC measures); 2) a focused literature review; and 3) TEP and expert input.

Candidate risk factors identified from work on related measures included opioid abuse, chronic anticoagulant use, tobacco use disorder, benign prostatic hyperplasia, morbid obesity, Work RVU, number of qualifying procedures, and procedure type. We used work RVU of the procedure to address surgical procedural complexity, an approach employed by the American College of Surgeons National Surgical Quality Improvement Program (NSQIP) [3].

To identify additional clinical and procedural risk factors, we searched the literature for relevant peer-reviewed publications of variables that predicted hospital visits after outpatient general surgery procedures using Ovid MEDLINE® and PubMed. The search yielded a total of 138 studies potentially relevant to the general surgery measure. Of these studies, 131 were excluded after review of the abstract, and 3 were excluded after full-text review. We added variables identified in the literature to our list of candidate risk factors if they were significantly associated with unplanned hospital visits in bivariate or multivariable analyses at the 0.05 level. From the 4 studies, we identified two variables not already included: anesthesia type and operating time [4-5]. However, we did not include anesthesia type or operating time because we do not risk adjust for discretionary procedure differences (such as approach to anesthesia or surgical techniques).

To define the clinical risk factors in claims data, we used CMS's Version 22 Hierarchical Condition Categories (HCCs) to operationalize the candidate clinical comorbidities. The HCCs classify 68,000 ICD-10-CM and over 15,000 ICD-9-CM diagnosis codes into clinically coherent condition categories. Then, to consolidate similar risk factors into fewer, broader risk variables, we first examined their frequency, bivariate direction and strength of association with the outcome of the individual risk factors defined by condition categories or ICD-10-CM codes, and then combined risk factor diagnoses into clinically coherent comorbidity variables. For example, we created a "cancer" variable that combined several individual cancer diagnoses.

Our expert clinical consultants and the TEP reviewed this preliminary list of risk variables and suggested additional variables: failure to thrive (poor nutritional status), history of falling, sleep apnea, and history of steroid use. We added all suggested candidate variables; the final list included 80 candidate risk variables.

Variable Selection

To select the final set of variables to include in the risk-adjustment model, risk variables were entered into logistic regression analyses predicting the outcome of hospital visits within 7 days in the Development Sample. The Development Sample is a randomly selected 50% sample of our CY 2015 Medicare cohort. To develop a parsimonious risk model, non-significant variables were iteratively removed from the model using a stepwise, purposeful selection approach described by Hosmer and Lemeshow [6]. All variables significant at $p<0.05$ were retained in the final model. The attached Data Dictionary sheet labeled "2b4.3 Risk Model Specs" indicates the final risk variables selected, the codes used to define the risk variables for our statistical model, and their odds ratios and 95% confidence intervals (CIs).

Social Risk Factors for supplementary disparities analyses

We selected variables representing SES factors and race based on a review of literature, conceptual pathways, and feasibility. In Section 1.8, we describe the variables that we considered and analyzed based on this review. Below, we describe the pathways by which SES and race may influence risk of hospital visits following outpatient surgical procedures.

Our conceptualization of the pathways by which patient SES or race affects the outcome is informed by the literature [7-12] and IMPACT Act–funded work by the National Academies of Sciences, Engineering and Medicine (NASEM) and the Department of Health and Human Services Assistant Secretary for Planning and Evaluation (ASPE) [13-15].

Literature Review of SES and Race Variables and Ambulatory Surgery Post-Procedure Hospital Visits

To examine the relationship between SES and race variables and risk of hospital visits following outpatient surgical procedures, a literature search was performed with the following exclusion criteria: non-English language articles, articles published more than 10 years ago, articles without primary data, articles focused on pediatric patient population, and articles not explicitly focused on SES or race and hospital visits after ambulatory surgery. A total of 176 studies were reviewed by title and abstract, and all but two studies were excluded from full-text review based on the above criteria. The two studies indicated that African-American and Hispanic patients and patients from lower-income households were at increased risk of post-procedure hospital visits in the ambulatory surgery setting [7-8]. No studies were found that suggested that variation in patients' SES and race affected variation in outcome risk across facilities performing ambulatory surgical procedures.

Conceptual Pathways for SES and Race Variable Selection

Although there is limited literature linking social risk factors and adverse outcomes, potential pathways may include:

1. **Differential care within an ASC or unmet differential needs.** One pathway by which SES factors or race may contribute to hospital visit risk is that patients may not receive equivalent care within a facility. In the hospital setting, African-American patients have been shown to experience differential, lower quality, or discriminatory care [9]. Alternatively, patients with SES risk factors, such as lower education, may require differentiated care – e.g., provision of information at a lower health literacy level – that they do not receive.
2. **Use of lower-quality facilities.** Patients may differentially obtain care in lower quality ASCs. With respect to hospital care, patients of lower income, lower education, or unstable housing have been shown not to have equitable access to high-quality facilities because such facilities are less likely to be found in geographic areas with large populations of poor patients. Thus, patients with low income are more likely to be seen in lower-quality hospitals, which can contribute to increased risk of adverse outcomes following hospitalization [10-11]. Similarly, African-American patients have been shown to have less access to high-quality hospitals compared with white patients [12]. It is unknown to what extent this may be true in the ambulatory surgery setting.
3. **Influence of SES on hospital visit risk outside of ASC quality.** Some SES risk factors, such as income or wealth, may affect the likelihood of post-procedure hospital visits without directly being associated with the quality of care received at the ASC. For instance, while an ASC may make appropriate care decisions and provide tailored care and education, a lower-income patient may have a worse outcome post-procedure due to a limited understanding of the discharge plan or a lack of home support, transportation or other resources for following it fully.

As indicated in Section 1.8, the SES and race variables that we examined are:

- Dual-eligible status
- African-American race
- AHRQ-validated SES index score

The description of the analyses related to social risk factors can be found in Section 2b3.4b below.

Citations

1. Krumholz HM, Brindis RG, Brush JE, et al. Standards for statistical models used for public reporting of health outcomes: An American Heart Association scientific statement from the Quality of Care and Outcomes Research Interdisciplinary Writing Group: Cosponsored by the Council on Epidemiology and Prevention and the Stroke Council endorsed by the American College of Cardiology Foundation. *Circulation*. 2006;113(3):456-462.

2. Normand S-LT, Shahian DM. Statistical and clinical aspects of hospital outcomes profiling. *Stat Sci*. 2007;22(2):206-226.

3. Raval MV, Cohen ME, Ingraham AM, et al. Improving American College of Surgeons National Surgical Quality Improvement Program risk adjustment: incorporation of a novel procedure risk score. *J Am Coll Surg.* 2010;211(6):715-723
4. Fleisher LA, Pasternak LR, Lyles A. A novel index of elevated risk of inpatient hospital admission immediately following outpatient surgery. *Arch Surg.* 2007;142(3):263-268.
5. Mioton LM, Buck DW, 2nd, Rambachan A, Ver Halen J, Dumanian GA, Kim JY. Predictors of readmission after outpatient plastic surgery. *Plast Reconstr Surg.* 2014;133(1):173-180.
6. Hosmer DW, Lemeshow S. Applied Logistic Regression. New York: Wiley; 2000.
7. Bhattacharyya N. Healthcare disparities in revisits for complications after adult tonsillectomy. *Am J Otolaryngol*. 2015 Mar-Apr;36(2):249-253.
8. Menachemi N, Chukmaitov A, Brown LS, et al. Quality of care differs by patient characteristics: outcome disparities after ambulatory surgical procedures. *Am J Med Qual*. 2007 Nov-Dec;22(6):395-401.
9. Trivedi AN, Nsa W, Hausmann LR, et al. Quality and equity of care in U.S. hospitals. *New Engl J Med*. 2014; 371:2298-2308.
10. Jha AK, Orav EJ, Epstein AM. Low-quality, high-cost hospitals, mainly in South, care for sharply higher shares of elderly black, Hispanic, and Medicaid patients. *Health Aff*. 2011; 30:1904-1911.
11. Reames BN, Birkmeyer NJ, Dimick JB, et al. Socioeconomic disparities in mortality after cancer surgery: failure to rescue. *JAMA Surg.* 2014; 149:475-481.
12. Skinner J, Chandra A, Staiger D, et al. Mortality after acute myocardial infarction in hospitals that disproportionately treat black patients. *Circulation* 2005; 112:2634-2641.

13. Department of Health and Human Services, Office of the Assistant Secretary of Planning and Evaluation. Report to Congress: Social Risk factors and Performance Under Medicare's Value-based Payment Programs. 2016; https://aspe.hhs.gov/pdf-report/report-congress-social-risk-factors-and-performance-under-medicares-value-based-purchasing-programs. Accessed November 10, 2017.

14. National Academies of Sciences, Engineering, and Medicine (NASEM); *Accounting for Social Risk Factors in Medicare Payment: Identifying Social Risk Factors.* Washington DC: National Academies Press; 2016.

15. National Academies of Sciences, Engineering, and Medicine (NASEM);. *Accounting for Social Risk Factors in Medicare Payment: Data.* Washington DC: National Academies Press; 2016.

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

☒ **Published literature**

☒ **Internal data analysis**

☐ **Other (please describe)**

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

The following candidate variables were significant at $p<0.05$ and were retained in the final model:
1. Age
2. Procedure Type: Abdomen and its contents
3. Procedure Type: Alimentary tract
4. Procedure Type: Breast
5. Procedure Type: Skin/soft tissue
6. Procedure Type: Wound
7. Procedure Type: Varicose vein

8. Work Relative Value Units
9. Other benign tumors (CC 15, 16)
10. Liver or biliary disease (CC 27, 28, 29, 30, 31)
11. Intestinal obstruction or perforation (CC 33)
12. Dementia or senility (CC 51, 52, 53)
13. Psychiatric disorders (CC 57, 58, 59, 60, 61, 62, 63)
14. Other significant central nervous system (CNS) disease (CC 77, 78, 79, 80)
15. Ischemic heart disease (CC 86, 87, 88, 89)
16. Specified arrhythmias and other heart rhythm disorders (CC 96, 97)
17. Stroke (CC 99, 100)
18. Chronic lung disease (CC 110, 111, 112, 113)
19. Pneumonia (CC 114, 115, 116)
20. Dialysis or sever chronic kidney disease (CC 134, 136, 137)
21. Benign prostatic hyperplasia (ICD-9 codes: 60000, 60001, 60020, 60021, 60090, 6091; ICD-10 codes: N40.0, N40.1, N40.2, N40.3)
22. Cellulitis, local skin infection (CC 164)
23. Major traumatic fracture or internal injury (CC 169, 170, 171, 172, 173, 174)
24. Complications of care (CC 176, 177)
25. Chronic anticoagulant use (ICD-9 code: V5861; ICD-10 code: Z7901 [long-term {current} use of anticoagulants])
26. Opioid abuse (ICD-9 codes: 30400, 30401, 30402, 30403, 30470, 30471, 30472, 30403, 30550, 30551, 30552, 30553; ICD-10: codes: F11.10, F11.120, F11.121, F11.122, F11.129, F11.14, F11.150, F11.151, F11.159, F11.181, F11.182, F11.188, F11.19, F11.20, F11.21, F11.220, F11.221, F11.222, F11.229, F11.23, F11.24, F11.250, F11.251, F11.259, F11.281, F11.282, F11.288, F11.29)
27. Procedure type*RVU

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

Methods

To examine the impact of social risk factors on the measure calculation, we evaluated three indicators of social risk: 1) Medicare-Medicaid dual eligibility, 2) race, and 3) the AHRQ SES Index. For these analyses we used 100% Medicare FFS claims data from CYs 2014-2015. These data included 3,653 ASC facilities and 303,220 general surgery procedures. Our goal for these analyses were twofold: 1) to examine whether these factors were associated with increased risk in hospital visits after adjusting for other risk factors and 2) to evaluate the impact of social risk factors on ASC-level measure scores.

To evaluate the association of these risk factors with the outcome, we first quantified the observed rate by each group (dual-eligible: yes vs. no, race: African-American vs. all others, AHRQ SES Index: lowest quartile of SES Index vs. all others). We next evaluated the magnitude of association of these social risk factors with the outcome after adjustment for clinical comorbidities, procedure type, and age by including each individual indicator as a variable in our risk-adjustment model. Each factor's effect was quantified using odds ratios (ORs) and tested for significance. In addition, we evaluated the change in each model's predictive ability (c-statistic).

To evaluate the impact of social risk factors on the ASC-level measure scores, we compared RSHVRs calculated with and without each disparity marker included in the model. For these analyses, we calculated the RSHVR

difference for each ASC (RSVHR with social risk variable – RSHVR without social risk variable) and calculated Pearson correlation coefficients for the paired scores.

We further examined the potential impact of these social risk factors on measure scores by comparing RSHVR distributions using current specifications. ASCs were stratified by the proportion of patients at the ASC with each factor, and placed into quartiles based on these proportions. For example, ASCs with few dual-eligible beneficiaries in their sample would be in the first quartile while ASCs seeing high numbers of dual-eligible beneficiaries would be in the fourth quartile. These stratified distributions were examined for systematic differences in RSHVR across quartiles.

Results

Observed hospital visit rates were higher for patients with each disparity marker: 3.7% for dual-eligible patients compared to 2.2% for non-dual-eligible patients, 3.1% for African-American patients compared to 2.2% for non-African-American patients, and 2.7% for low SES patients (scores below 42.7 on the AHRQ SES Index) compared to 2.2% for higher SES patients (scores above 42.7 on the AHRQ SES index). Furthermore, inclusion of each of these risk factors in our models indicated a statistically significant association after controlling for other risk-adjusters in our model (dual-eligible: OR: 1.34, 95% CI: 1.22 -1.48, $p < 0.0001$; race: OR: 1.23, 95% CI: 1.06-1.42, p=0.005; AHRQ SES Index: OR: 1.14, 95% CI: 1.06-1.22, p=0.0004).

However, results of examining the impact of social risk factors on the ASC-level measure scores indicated that entering these variables into the risk-adjustment model did not improve model performance (c-statistics remained unchanged) and did not substantially change ASC-level measure scores. Correlation coefficients between RSHVRs with and without adjustment for these factors were near 1 (0.998, 1.000, and 0.999 for dual-eligible, African-American, and low SES patients, respectively) and mean differences in RSHVRs were near zero (0.0000, -0.0001, and -0.0002 for dual-eligible, African-American, and low SES patients, respectively).

Further, the analyses of ASCs stratified into quartiles based on proportions of dual-eligible, African-American, and low SES patients (as identified by the AHRQ SES Index) showed largely overlapping distributions of the RSHVRs by quartile. The median RSHVR was 1.0 for all three variables except for ASCs with a low % of dual-eligible patients (1st quartile) whose median RSHVR was 0.9. Longer tails at the upper ends of the distributions were observed for ASCs with the highest percent of patients with the social risk factor (4th quartile). Distributions for low % of social risk factor ASCs (1st quartile) and high % social risk factor ASCs (4th quartile) by each social risk factor are shown in below in Table 2.

Based on these analyses, we conclude that although the three social risk factors we examined have a modest but statistically significant association with the risk of a hospital visit, these patient-level factors have a limited effect on the ASC-level measure scores. We did not adjust the models for these social risk factors since the association of these factors with the outcome may be quality-related, and since these factors have a limited relationship to the facility-level scores.

Table 2. Variation in RSHVRs across ASCs grouped into quartiles by proportion of Medicaid dual-eligible, African-American race, and Low SES patients

Social Risk Factor	Medicaid Dual Eligible	African-American Race	Low AHRQ SES Index Score

Proportion of ASC patients	1st Quartile (≤1.82%)	4th Quartile (≥7.06%)	1st Quartile (0%)	4th Quartile (≥3.95%)	1st Quartile (≤4.04%)	4th Quartile (≥17.17%)
# of ASCs	409	411	599	410	410	410
# of patients	83,214	48,895	79,947	48,318	71,841	63,945
100% Max	1.6	1.9	1.9	2.1	1.8	1.9
90%	1.2	1.3	1.2	1.3	1.2	1.3
75% Q3	1.1	1.1	1.1	1.1	1.1	1.1
50% Median	0.9	1.0	1.0	1.0	1.0	1.0
25% Q1	0.9	0.9	0.9	0.9	0.9	0.9
10%	0.8	0.8	0.8	0.9	0.8	0.8
0% Min	0.5	0.4	0.6	0.4	0.5	0.6

2b3.5. Describe the method of testing/analysis used to develop and validate the adequacy of the statistical model or stratification approach (*describe the steps—do not just name a method; what statistical analysis was used*)
Provide the statistical results from testing the approach to controlling for differences in patient characteristics (case mix) below.
If stratified, skip to 2b3.9

To assess performance of the patient-level risk-adjustment model in the Development Sample, the area under the receiver operating characteristic curve as measured by the c-statistic was calculated. Observed hospital visit rates were compared to predicted hospital visit probabilities across predicted risk deciles to assess calibration, and the range of observed hospital visit rates between the lowest and highest predicted risk deciles was also calculated to assess model discrimination.

Several analyses to validate the patient-level risk-adjustment model were performed. First, we compared model performance in the Development Sample with its performance in the Validation Sample. The c-statistic, and model predictive ability (discrimination) were compared. Second, we examined the stability of the risk variable frequencies and regression coefficients across the Development and Validation Samples. Third, we calculated over-fitting indices in the Validation Sample. Over-fitting refers to the phenomenon in which a model describes the relationship between predictive variables and outcome well in the development dataset but fails to provide valid predictions in new patients in the validation dataset. Estimated calibration values of $\gamma 0$ far from 0 and estimated values of $\gamma 1$ far from 1 provide evidence of over-fitting.

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

Development Sample results:

c-statistic=0.699
Predictive ability (hospital visit % in lowest decile, hospital visit % in highest decile): 0.79%-6.39%

Validation Sample results:

c-statistic=0.700
Predictive ability (hospital visit % in lowest decile, hospital visit % in highest decile): 0.71%-6.44%

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

Development Sample results:

Calibration: (0, 1)

Validation Sample results:

Calibration: (-0.08, 0.98)

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

Below are plots of observed vs. predicted values for the hospital visit rate across deciles of patient risk in the Development Sample (Figure 1) and Validation Sample (Figure 2). The plots, which showed that the predicted risk closely approximated the observed risk in most deciles, suggest reasonable calibration.

Figure 1. Calibration plot of predicted versus observed outcomes across deciles of patient risk in the Development Sample (data source: Medicare FFS CY 2015 dataset)

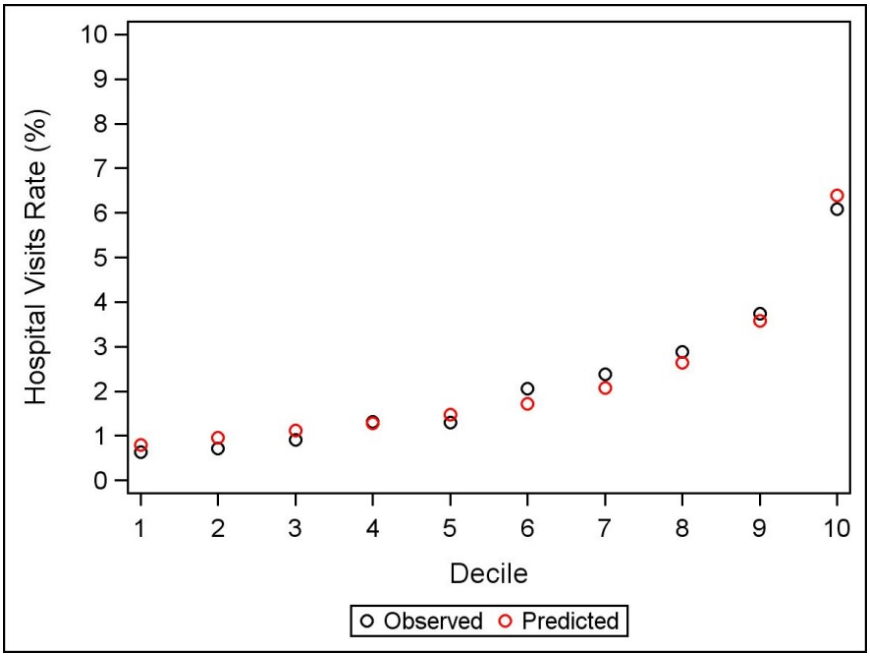
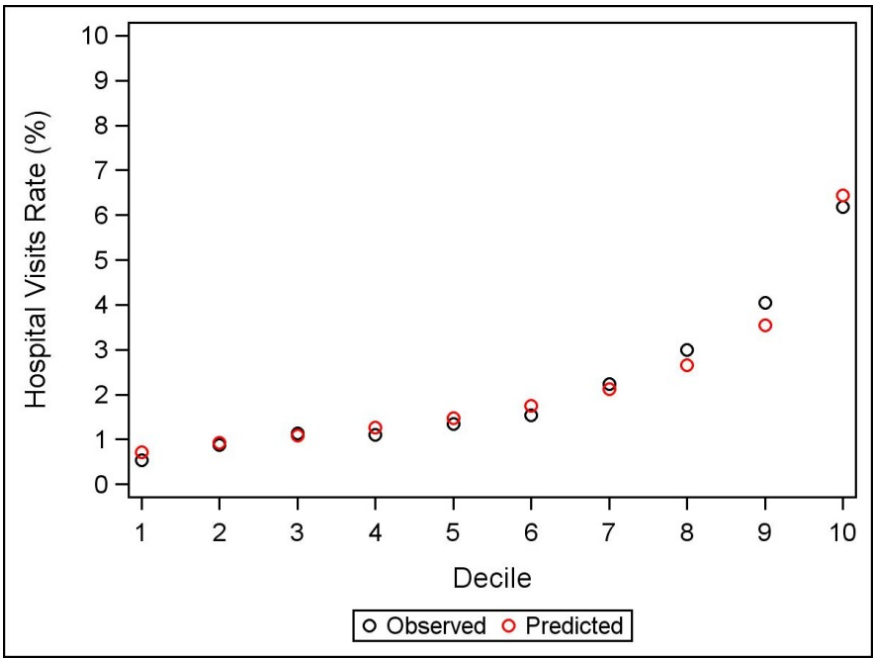


Figure 2. Calibration plot of predicted versus observed outcomes across deciles of patient risk in the 2015 Validation Sample (data source: Medicare FFS CY 2015 dataset)



2b3.9. Results of Risk Stratification Analysis:

Not applicable. This measure is not risk-stratified.

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

The c-statistics in the Development Sample and the Validation Sample were 0.699 and 0.700, respectively, showing good discrimination. The risk decile plots, which showed that the predicted risk closely approximated the observed risk across deciles, suggest good model calibration. The predicted unplanned hospital visit rate in the Development Sample ranged from 0.79% in the lowest decile to 6.39% in the highest predicted risk decile, a range of 5.6%; comparable results were found in the Validation Sample. In addition, the regression coefficients of the model variables were stable across the Development and Validation Samples.

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

We tested interaction terms and retained those that were both significant at $p<0.01$ and clinically correlated with the outcome.

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

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

The measure score is an ASC-level RSHVR. The RSHVR is calculated as the ratio of the predicted to the expected number of post-surgical unplanned hospital visits among an ASC's patients. For each ASC, the numerator of the ratio is the number of hospital visits predicted for the ASC's patients, accounting for its observed rate, the number and complexity of general surgery procedures performed at the ASC, and the patient mix. The denominator is the number of hospital visits expected nationally for the ASC's case/procedure mix. To calculate an ASC's predicted-to-expected (P/E) ratio, the measure uses a two-level hierarchical logistic regression model. The log-odds of the outcome for an index procedure is modeled as a function of the patient demographic, comorbidity, procedure characteristics, and a random ASC-specific intercept. A ratio >1 indicates that the ASC's patients have more hospital visits than expected, compared to an average ASC with similar patient and procedural complexity. A ratio <1 indicates that the ASC's patients have fewer post-surgical visits than expected, compared to an average ASC with similar patient and procedural complexity.

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

The risk-standardized measure scores estimated using two full years of Medicare FFS data (2014 and 2015) showed variation across ASCs (Range: Min 0.42 to Max 2.13).

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

These results suggest there are meaningful differences in the quality of care provided to patients undergoing general surgery procedures at ASCs.

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

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

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

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

Items 2b5.1-2b5.3 are not applicable, as this measure has only one set of specifications.

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

Items 2b5.1-2b5.3 are not applicable, as this measure has only one set of specifications.

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

Items 2b5.1-2b5.3 are not applicable, as this measure has only one set of specifications.

2b6. MISSING DATA ANALYSIS AND MINIMIZING BIAS

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

Not applicable.

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

Not applicable.

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

Not applicable.

3. Feasibility

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

3a. Byproduct of Care Processes

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

3a.1. **Data Elements Generated as Byproduct of Care Processes.**
Coded by someone other than person obtaining original information (e.g., DRG, ICD-9 codes on claims)
If other:

3b. Electronic Sources

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

3b.1. **To what extent are the specified data elements available electronically in defined fields** (*i.e., data elements that are needed to compute the performance measure score are in defined, computer-readable fields*) Update this field for **maintenance of endorsement**.
No data elements are in defined fields in electronic sources

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

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

3c. Data Collection Strategy

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

3c.1. **Required for maintenance of endorsement. Describe difficulties (as a result of testing and/or operational use of the measure) regarding data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, time and cost of data collection, other feasibility/implementation issues.**
IF instrument-based, consider implications for both individuals providing data (patients, service recipients, respondents) and those whose performance is being measured.
Not applicable. Measure is not currently in use. However, measure development and testing show that the measure cohort can be defined and outcomes can be reported using routinely collected Medicare claims data.

3c.2. **Describe any fees, licensing, or other requirements to use any aspect of the measure as specified** *(e.g., value/code set, risk model, programming code, algorithm)*.
Not applicable. There are no fees, licensing, or other requirements to use any aspect of the measure as specified.

4. Usability and Use

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

4a. Accountability and Transparency

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

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

Specific Plan for Use	Current Use (for current use provide URL)
Payment Program Not in use	

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

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

Not applicable. Measure is not yet in use.

4a1.2. **If not currently publicly reported OR used in at least one other accountability application (e.g., payment program, certification, licensing) what are the reasons?** (*e.g., Do policies or actions of the developer/steward or accountable entities restrict access to performance results or impede implementation?*)
This measure is not currently publicly reported or used in an accountability application because the measure is still under development and is now being submitted to the National Quality Forum (NQF) for initial endorsement.

4a1.3. **If not currently publicly reported OR used in at least one other accountability application, provide a credible plan for implementation within the expected timeframes -- any accountability application within 3 years and publicly reported within 6 years of initial endorsement.** (*Credible plan includes the specific program, purpose, intended audience, and timeline for implementing the measure within the specified timeframes. A plan for accountability applications addresses mechanisms for data aggregation and reporting.*)
The measure may ultimately be used in one or more CMS programs, such as the Ambulatory Surgical Center Quality Reporting program.

4a2.1.1. **Describe how performance results, data, and assistance with interpretation have been provided to those being measured or other users during development or implementation.**
How many and which types of measured entities and/or others were included? If only a sample of measured entities were included, describe the full population and how the sample was selected.
During development of the measure, we recruited and met with a national TEP, and CMS hosted a public comment. CMS solicited public comments on the measure, and we took all comments into consideration, addressing them individually. Therefore, performance results and data were provided to members of the TEP and then made public through public comment. TEP members and commenters included representatives of the measured entities (ASCs). The exact number of measured entities (ASCs) varies with each measurement period. In the Medicare FFS 2015 Dataset we used for measure development, there were 149,468 general surgery procedures performed at 3,251 ASCs. In the Medicare FFS CYs 2014-2015 Dataset we used for calculating ASC-level measures, there were 286,999 general surgery procedures performed at 1,642 ASCs (with at least 25 cases). (See section 1.7 of Measure Testing Form for a complete description of the number of measure entities).

4a2.1.2. **Describe the process(es) involved, including when/how often results were provided, what data were provided, what educational/explanatory efforts were made, etc.**
We provided data and results to the TEP and obtained TEP input on five occasions throughout the measure development process. We hosted two teleconference meetings with the TEP, solicited TEP input via email on the risk model, and provided measure updates to the TEP twice via email in response to public comments we received on the measure.

4a2.2.1. **Summarize the feedback on measure performance and implementation from the measured entities and others described in 4d.1.**
Describe how feedback was obtained.
Not applicable; the measure has not yet been implemented. Feedback during development as obtained through a TEP and public comment as described in 4a2.1.1.

4a2.2.2. **Summarize the feedback obtained from those being measured.**
Not applicable; the measure has not yet been implemented. See section 4a2.3 below for how TEP and public comment feedback was considered during measure development.

4a2.2.3. **Summarize the feedback obtained from other users**
Not applicable; the measure has not yet been implemented.

4a2.3. **Describe how the feedback described in 4a2.2.1 has been considered when developing or revising the measure specifications or implementation, including whether the measure was modified and why or why not.**
This measure was developed with input from national TEP consisting of patients, surgeons, methodologists, researchers, and providers. We also held a three-week public comment period soliciting stakeholder input on the measure methodology, and publicly posted a summary of the comments received as well as our responses (available in the Downloads section at https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/MMS/PC-Updates-on-Previous-Comment-Periods.html.

CMS and The Center for Outcomes Research and Evaluation (CORE) investigated issues identified during measure development public comment. Specifically, CORE and CMS:

- Renamed the measure “Facility-Level 7-Day Hospital Visits after General Surgery Procedures Performed at Ambulatory Surgical Centers” to reflect the procedures included in the measure cohort.

- In response to feedback received during the measure development public comment period, reviewed all of the individual CPT® codes within CCS categories and removed 15 individual procedures (CPT® codes) from the measure that are outside the scope of general surgery practice, including two specifically suggested for removal by a commenter.

- Reviewed statistically selected variables for face validity for the final risk model, and retained one variable (opioid use) because experts advised it was an important risk predictor and expressed a strong preference for including it in the model even though it was not statistically selected.

Improvement

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

4b1. **Refer to data provided in 1b but do not repeat here. Discuss any progress on improvement (trends in performance results, number and percentage of people receiving high-quality healthcare; Geographic area and number and percentage of accountable entities and patients included.)**
If no improvement was demonstrated, what are the reasons? If not in use for performance improvement at the time of initial endorsement, provide a credible rationale that describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.
Not applicable. Measure is not yet in use.

4b2. Unintended Consequences

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

4b2.1. **Please explain any unexpected findings (positive or negative) during implementation of this measure including unintended impacts on patients.**
Not applicable. Measure is not yet in use.

4b2.2. **Please explain any unexpected benefits from implementation of this measure.**
Not applicable. Measure is not yet in use.

5. Comparison to Related or Competing Measures

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

5. Relation to Other NQF-endorsed Measures
Are there related measures (conceptually, either same measure focus or target population) or competing measures (conceptually both the same measure focus and same target population)? If yes, list the NQF # and title of all related and/or competing measures.
Yes

5.1a. **List of related or competing measures (selected from NQF-endorsed measures)**
2539 : Facility 7-Day Risk-Standardized Hospital Visit Rate after Outpatient Colonoscopy
2687 : Hospital Visits after Hospital Outpatient Surgery

5.1b. **If related or competing measures are not NQF endorsed please indicate measure title and steward.**
Not yet submitted to NQF: Hospital Visits after ASC Orthopedic Procedures (CMS)
Submitting to NQF in this November 2017 round: Hospital Visits after ASC Urology Procedures (CMS)

5a. Harmonization of Related Measures

The measure specifications are harmonized with related measures;
OR
The differences in specifications are justified

5a.1. **If this measure conceptually addresses EITHER the same measure focus OR the same target population as NQF-endorsed measure(s):**
Are the measure specifications harmonized to the extent possible?
Yes

5a.2. **If the measure specifications are not completely harmonized, identify the differences, rationale, and impact on interpretability and data collection burden.**
Not applicable. The measures' outcomes are harmonized.

5b. Competing Measures

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

5b.1. **If this measure conceptually addresses both the same measure focus and the same target population as NQF-endorsed measure(s):**
Describe why this measure is superior to competing measures (e.g., a more valid or efficient way to measure quality); OR provide a rationale for the additive value of endorsing an additional measure. (Provide analyses when possible.)
Not applicable. There are no competing measures.

Appendix

A.1 Supplemental materials may be provided in an appendix. All supplemental materials (such as data collection instrument or methodology reports) should be organized in one file with a table of contents or bookmarks. If material pertains to a specific submission form number, that should be indicated. Requested information should be provided in the submission form and required attachments. There is no guarantee that supplemental materials will be reviewed.
Attachment **Attachment:** Gen_Surg_ASC_NQF_Appendix_v2.0_-1-.pdf

Contact Information

Co.1 **Measure Steward (Intellectual Property Owner):** Centers for Medicare & Medicaid Services (CMS)
Co.2 **Point of Contact:** Vinitha, Meyyur, Vinitha.Meyyur@cms.hhs.gov, 410-786-8819-
Co.3 **Measure Developer if different from Measure Steward:** YNNH/Yale Center for Outcomes Research and Evaluation
Co.4 **Point of Contact:** Danielle, Purvis, Danielle.purvis@yale.edu, 203-200-5342-

Additional Information

Ad.1 **Workgroup/Expert Panel involved in measure development**
Provide a list of sponsoring organizations and workgroup/panel members' names and organizations. Describe the members' role in measure development.
CORE convened a TEP comprised of clinicians, patients, and experts in quality improvement to provide input on key methodological decisions.

TEP Members
- Robin Blomberg, BA, MA – National Forum of End-Stage Renal Disease, Network 16 (Representative for Kidney Patient Advisory Council); Seattle, WA
- Kirk Campbell, MD – New York University Hospital for Joint Diseases (Clinical Assistant Professor of Orthopedic Surgery); New York, NY
- Gary Culbertson, MD, FACS – Iris Surgery Center (Surgeon; Medical Director); Sumter, SC
- Martha Deed, PhD – Consumers Union Safe Patient Project (Patient Safety Advocate); Austin, TX
- James Dupree, MD, MPH – University of Michigan (Urologist; Health Services Researcher); Ann Arbor, MI
- Nester Esnaola, MD, MPH, MBA – Fox Chase Cancer Center (Professor of Surgery; Associate Director for Cancer Health Disparities and Community Engagement); Philadelphia, PA
- John Gore, MD, MS – University of Washington (Associate Professor of Urology); Seattle, WA
- Lisa Ishii, MD, MHS – Johns Hopkins School of Medicine (Associate Professor); American Academy of Otolaryngology-Head and Neck Surgery (Coordinator for Research and Quality); Baltimore, MD; Alexandria, VA
- Atul Kamath, MD – Perelman School of Medicine, University of Pennsylvania (Assistant Professor and Clinical Educator Director of Orthopedic Surgery); Hospital of the University of Pennsylvania (Attending Surgeon); Philadelphia, PA
- Tricia Meyer, PharmD, MS, FASHP – Scott & White Medical Center (Regional Director of Pharmacy); Texas A&M University College of Medicine (Associate Professor of Anesthesiology); Temple, TX
- Linda Radach, BA – Consumers Union Safe Patient Project (Patient Safety Advocate); Austin, TX
- Amita Rastogi, MD, MHA, CHE, MS – Health Care Incentives Improvement Institute (Chief Medical Officer); Newtown, CT
- Donna Slosburg, RN, BSN, LHRM, CASC – ASC Quality Collaboration (Executive Director); St. Pete Beach, FL
- Julie Thacker, MD, FACS – Duke Health and Hospital System (Medical Director of Evidence-Based Perioperative Care); Duke School of Medicine Clinical Research Unit (Medical Director, Department of Surgery); Durham, NC

- Thomas Tsai, MD, MPH – Brigham and Women's Hospital (General Surgeon); Harvard School of Public Health (Research Associate); Boston, MA

The CORE measure development team met regularly and was comprised of experts in internal medicine, quality outcomes measurement, and measure development. CORE convened surgical and statistical consultants with expertise relevant to outpatient surgery and quality measurement to provide input on key methodological decisions.

CORE Measure Development Team
- Faseeha Altaf, MPH – Project Lead, CORE
- Haikun Bao, PhD – Analytic Lead, CORE
- Mayur Desai, PhD, MPH – Project Consultant, CORE
- Elizabeth Drye, MD, SM – Project Director, CORE
- Harlan Krumholz, MD, SM – Director, CORE
- Zhenqiu Lin, PhD – Analytics Director, CORE
- Megan LoDolce, MA – Project Manager, CORE
- Erica Norton, BS – Research Associate, CORE
- Danielle Purvis, MPH – Project Coordinator, CORE
- Craig Parzynski, MS – Analytic Consultant, CORE
- Jennifer Schwartz, PhD, MPH – Project Lead (Formerly at CORE)
- Rushi Shah, BS – Research Assistant, CORE
- Mona Sharifi, MD, MPH – Clinical Consultant, Instructor of Pediatrics, Yale University School of Medicine

Consultants
- Robert Becher, MD, MS – Surgical Consultant, Assistant Professor of Surgery at Yale University School of Medicine
- Sean O'Neill, ND, PhD – University of California, Los Angeles (Resident, General surgery); Los Angeles, CA
- Sharon-Lise Normand, PhD, MSc—Statistical Consultant, Professor of Biostatistics, Department of Health Care Policy, Harvard Medical School

Measure Developer/Steward Updates and Ongoing Maintenance

Ad.2 **Year the measure was first released:**

Ad.3 **Month and Year of most recent revision:**

Ad.4 **What is your frequency for review/update of this measure?** Not applicable; not yet endorsed

Ad.5 **When is the next scheduled review/update for this measure?**

Ad.6 **Copyright statement:** Not applicable.

Ad.7 **Disclaimers:** Not applicable.

Ad.8 **Additional Information/Comments:** None.

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
Measure Title: STS Lobectomy for Lung Cancer Composite Score
Measure Steward: The Society of Thoracic Surgeons
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
Developer Rationale: n/a

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.

Denominator Statement: Number of patients greater than or equal to 18 years of age undergoing elective lobectomy for lung cancer

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
Data Source: Other, Registry Data
Level of Analysis: Facility

IF Endorsement Maintenance – Original Endorsement Date: Most Recent Endorsement Date:

New Measure – Preliminary Analysis

Criteria 1: Importance to Measure and Report

1a. Evidence

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.

Evidence Summary

- This new measure assesses the operative mortality and the presence of at least one of 9 major complications of lobectomy, the most frequently performed lung resection procedure. The developer reports that data in the STS General Thoracic Surgery database (GTSD) show a reduction in perioperative morbidity and equivalent long term survival when minimally invasive approaches for lobectomy are used.
- The developer provided the performance data below, 0.95 to 0.98, for approximately 200-300 participants and 24,000+ operations from 2013 to 2016.
- *Empirical data* demonstrating a relationship between the outcome to at least one healthcare process is now required. NQF guidance states that a wide variation in performance can be used as evidence, assuming the data are from a robust number of providers and results are not subject to systematic bias.

Question for the Committee:

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

Guidance from the Evidence Algorithm: Measure assesses performance on a health outcome (Box 1) → The relationship between the outcome and the intervention demonstrated by demonstrated by performance data (Box 2) → Pass

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

1b. Gap in Care/Opportunity for Improvement and 1b. Disparities
Maintenance measures – increased emphasis on gap and variation

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

- Data were collected in two overlapping 3 year time periods: January 1, 2014 – December 31, 2016 and January 1, 2013 – December 31, 2015.

January 1, 2013 – December 31, 2015			January 1, 2014 – December 31, 2016	
No. participants	242	185	233	286
No. of operations	23,574	22,572	24,912	24,318
Mean	0.972	0.972	0.973	0.974
Standard Deviation	0.007	0.008	0.006	0.007
IQR	0.008	0.009	0.007	0.009
Minimum	0.945	0.945	0.953	0.953
Maximum	0.988	0.988	0.987	0.987

Disparities

- The developer provides descriptive data of the sampled population, but disparities data for these groups are not provided.

Questions for the Committee:

o *Does the Committee think there is enough variation among providers to justify a national performance measure?*

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.

Quality construct

- This measure is based on a combination of an operative mortality outcome and the risk adjusted occurrence of any of nine major complications. 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
    - Branchopleural 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 difficult to differentiate performance based on mortality alone since it fails to take into account 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. The developer reports this weighting is consistent with STS adult cardiac measures.

Questions for the Committee:

o *Are the quality construct and a rationale for the composite explicitly stated and logical?*

o *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

Criteria 1: Importance to Measure and Report (including 1a, 1b, 1c)

1a. Evidence
**STS General Thoracic Surgery database (GTSD) 200-300 patients 24,000 patients (?) PASS
**good evidence
**Adequate evidence
**An important measure for public accountability, as already illustrated by improvement in outcomes over the course of the registry. Weighting death 4x morbidities is somewhat arbitrary, but reasonable and consistent with other such measures.

1b. Performance Gap
**Improvement, Two year data, MODERATE per NQF reviewer "
**PG present
**Increasing morbidity associated with lobectomy clearly justifies this composite measure
**Minimal gap (91% average performance), so limited opportunity for quality improvement. But as noted above, important for public accountability.

1c. Composite Quality Construct
**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. Logical
**High quality composite construct
**The construct makes good sense and adds value to the individual components
**See above. Well constructed statistically, but balance between different outcomes will always be arbitrary.

Criteria 2: Scientific Acceptability of Measure Properties
2a. Reliability: Specifications and Testing
2b. Validity: Testing; Exclusions; Risk-Adjustment; Meaningful Differences; Comparability Missing Data
2c. For composite measures: empirical analysis support composite approach

Reliability
2a1. Specifications requires the measure, as specified, to produce consistent (reliable) and credible (valid) results about the quality of care when implemented. For maintenance measures – no change in emphasis – specifications should be evaluated the same as with new measures.
2a2. Reliability testing demonstrates if the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise enough to distinguish differences in performance across providers. For maintenance measures – less emphasis if no new testing data provided.
Validity
2b2. Validity testing should demonstrate the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality. For maintenance measures – less emphasis if no new testing data provided.
2b2-2b6. Potential threats to validity should be assessed/addressed.
Composite measures only:
2d. Empirical analysis to support composite construction. Empirical analysis should demonstrate that the component measures add value to the composite and that the aggregation and weighting rules are consistent with the quality construct.

Complex measure evaluated by Scientific Methods Panel? ☒ **Yes** ☐ **No**
Evaluators: Jennifer Perloff, Ron Walters, Joe Kunisch, David Cella, Karen Maddox

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

Evaluation A
Evaluation B
Evaluation C
Evaluation D
Evaluation E

Questions for the Committee regarding reliability:

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

Questions for the Committee regarding validity:

- *Do you have any concerns regarding the validity of the measure (e.g., exclusions, risk-adjustment approach, etc.)?*
- *The Scientific Methods Panel is satisfied with the validity analyses for the measure. Does the Committee think there is a need to discuss and/or vote on validity?*

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?)?*
- *The Scientific Methods Panel is satisfied with the composite construction. Does the Committee think there is a need to discuss the composite construction approach?*

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

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

Note: While score-level validity testing is desired, data element testing is accepted because this is a new measure. For future maintenance evaluations, score-level testing will be required.

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

Committee pre-evaluation comments

Criteria 2: Scientific Acceptability of Measure Properties (including all 2a, 2b, and 2c)

2a1. Reliability specifications
**STS database. No concerns about implementation.
**Reliable
**Well-defined
**No issues.

2a2. Reliability testing
**STS database. No
**No
**No concerns
**No issues

2b1. Validity Testing

**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.
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 no threat to validity
**No concerns
**Data abstracted from clinical records - minimal concerns re: data validity
**No issues

2b2.-3. Other threats to validity
**Risk adjustment rigorous
**Adequate – the usual problem with random effects models of squishing outcomes towards the mean, especially for low-volume groups

2c. Composite Analysis
****Fits quality construct and rationale
STS's combined mortality and morbidity model for pulmonary resection for lung cancer is important and appropriate for public reporting for the following reasons:
1.) within the broad category of lung cancer resections, lobectomy is the single most common major procedure that a thoracic surgeon performs;
2.) These procedures are therefore useful and appropriate to use as a benchmark for performance by general thoracic surgery programs. By providing surgeons and teams with risk-adjusted results, they can identify how they are performing compared with other programs in the STS General Thoracic Database,
Most recently, STS surgeon members have expressed interest in real-time, online data updates, which has led to the development of dashboard-type reporting on STS.org. The general thoracic dashboard is scheduled for launch in 2018.
**Composite measure credibly reflects pt. experience
**Yes and yes
**No issues.

Criterion 3. Feasibility

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

- The developer reports that data are generated or collected by and used by healthcare personnel during the provision of care; coded by someone other than the person obtaining original information; and abstracted from a record by someone other than the person obtaining the original information.
- All data are in defined fields in a combination of electronic sources
- Data are collected continuously by the participating sites and harvested by DCRI twice a year; reports are then sent back to participating sites about three months after harvest. Participating sites generally have data managers on staff.
- The developer reports that STS GTSD participant surgeons pay an annual participant fee of $550 or $700 depending on whether the participant is an STS member.

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., EHR or other electronic sources?*
- *Is the data collection strategy ready to be put into operational use?*

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

Committee pre-evaluation comments

Criteria 3: Feasibility

3. Feasibility
**STS database feasibility good

**feasible
**Feasible through the GTSD
**Requires participation in the registry -- impossible to replicate/participate otherwise.

Criterion 4: Usability and Use

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

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

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

Current uses of the measure
Publicly reported? ☒ **Yes** ☐ **No**

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

Accountability program details

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

4a.2. Feedback on the measure by those being measured or others. Three criteria demonstrate feedback: 1) those being measured have been given performance results or data, as well as assistance with interpreting the measure results and data; 2) those being measured and other users have been given an opportunity to provide feedback on the measure performance or implementation; 3) this feedback has been considered when changes are incorporated into the measure

Feedback on the measure by those being measured or others

- The developer states that STS surgeon members have expressed interest in real-time, online data updates which led to the development of a general thoracic dashboard. The dashboard is scheduled for launch in 2018.
- The developer states that given the recent launch of public reporting that they have not received sufficient feedback from non-participants to be able to assess the impact of the public reporting initiative.

Additional Feedback:

- The developer reports that surgeons on the STS General Thoracic Surgery Task Force meet periodically to discuss participant reports and discuss enhancements to the GTS database. Additions and clarifications to the data collection form and the content/format of participant reports are discussed and implemented as appropriate.

***Questions for the Committee**:*

- *How 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 (4a1. Improvement; 4a2. Benefits of measure)

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

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

Improvement results

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

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

Unexpected findings (positive or negative) during implementation

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

Potential harms

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

***Questions for the Committee**:*

- *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

Criteria 4: Usability and Use

4a1. Use

**Most recently, STS surgeon members have expressed interest in real-time, online data updates, which has led to the development of dashboard-type reporting on STS.org. The general thoracic dashboard is scheduled for launch in 2018. Star ratings for surgeons and hospitals will be developed
**usable
**Measure is not being used in an accountability program but is being publicly reported
**Already publicly reported

4b1. Usability

**Believe the benefits outweigh unintended consequences. Recommend Approval
**No concerns. Separately I am worried about additive value of this measure compared to measure 1790
**Overall benefits outweigh harms
**Yes

Criterion 5: Related and Competing Measures

Related or competing measures

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

Harmonization

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

Committee pre-evaluation comments

Criterion 5: Related and Competing Measures

Public and member comments

Comments and Member Support/Non-Support Submitted as of: January 23, 2018

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

Evaluation A

Scientific Acceptability

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

Instructions:

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

Measure Number: NQF#3294
Measure Title: STS Lobectomy for Lung Cancer Composite Score

RELIABILITY

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

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

   ☒Yes (go to Question #2)

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

   The measure clearly defines the intended outcomes measures, mortality and complications of care. These have been utilized by this database for many years and are consistent with prior definitions in previous implementations. The measure is intended at the facility level AND the individual group/practice level, however, data is provided for 233 participants (facilities) and 24,912 patient records. The data elements are clearly defined and annually audited for data completeness and accuracy.

   Data presented, however, seems to be at the facility level (233/24,912) and I could not find comparable testing at the group/practice level from the submission. The term participant in both the submission and the publication referenced appears to apply to the facility level only.

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

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

   ☒Yes (go to Question #4)

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

   Reliability at the data element level (in 2a2.4) of 44.6% for all and higher for increasing number of cases (up to 68.0%) and the score level (one star to three stars) with the weighted composite is tested and indicated in 2d1.2.

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

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

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

   Data element validity testing is stated in 1.7 as being via an annual audit of data completeness and accuracy for randomly selected surgical records at randomly selected participant sites, described in 2b1.2. A data element quality report is generated and provided to the participant for action, if required. Agreement was 97.78% in 2016.

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

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

   ☒Yes (go to Question #5)

   ☐No (go to Question #8)

   Yes at the facility level. No at the group/practice level.

5. Was the method described and appropriate for assessing the proportion of variability due to real differences among measured entities? *NOTE: If multiple methods used, at least one must be appropriate.*
   TIPS: Examples of appropriate methods include signal-to-noise analysis (e.g. Adams/RAND tutorial); random split-half correlation; other accepted method with description of how it assesses reliability of the performance score.
   ☒Yes (go to Question #6)
   ☐No (please explain below then go to Question #8)

   Section 2a2.2 provides the metholodogy and the results.
6. **RATING (score level)** - What is the level of certainty or confidence that the performance measure scores are reliable?
   TIPS: Consider the following: Is the test sample adequate to generalize for widespread implementation?
   Do the results demonstrate sufficient reliability so that differences in performance can be identified?
   ☒High (go to Question #8)
   ☐Moderate (go to Question #8)
   ☐Low (please explain below then go to Question #7)

   At the facility level, score reliability does separate out those with high mortality, 1.2% and complications, 16.2% (one star) from those with low mortality, 0.4% and complications, 3.2%, (three star). An expert panel provided an assessment of validity. The methodology is described in 2b4.1.

7. Was other reliability testing reported?
   ☐Yes (go to Question #8)
   ☐No (rate Question #11: OVERALL RELIABILITY as LOW and proceed to the VALIDITY SECTION)

8. Was reliability testing conducted with patient-level data elements that are used to construct the performance measure?
   TIPS: Prior reliability studies of the same data elements may be submitted; if comparing abstraction to "authoritative source/gold standard" see Validity Section Question #15)
   ☐Yes (go to Question #9)
   ☒No (if there is score-level testing, rate Question #11: OVERALL RELIABILITY based on score-level rating from Question #6; otherwise, rate Question #11: OVERALL RELIABILITY as INSUFFICIENT. Then proceed to the VALIDITY SECTION)

   The inter-rater or intra-rater reliability testing is not specifically given in this measure submission but I suspect is known from prior experience with this dataset. The referenced data in the submission is to the audits for data accuracy. Thus, though I suspect the answer to this question is YES, I cannot state this from the data given.

9. Was the method described and appropriate for assessing the reliability of ALL critical data elements?
   TIPS: For example: inter-abstractor agreement (ICC, Kappa); other accepted method with description of how it assesses reliability of the data elements
   Answer no if: only assessed percent agreement; did not assess separately for all data elements (at least numerator, denominator, exclusions)
   ☐Yes (go to Question #10)
   ☐No (if no, please explain below and rate Question #10 as INSUFFICIENT)

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

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

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

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

    ☐Insufficient (go to Question #11)

11. OVERALL RELIABILITY RATING

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

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

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

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

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

Information is not provided about the inter-rater reliability at the data element level and is substituted by the results of random audits of data elements.

If, by the term “participant”, both facility and group/practice level is the intention and has been performed (see Question 1), and if there is prior evidence of inter-rater and intra-rater reliability testing historically, not based on random audits, then I would be willing to consider changing the overall rating to high. I could not infer this from the submission and these points should be clarified and discussed.

VALIDITY

Assessment of Threats to Validity

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

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

   ☒Yes (go to Question #2)

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

   There was acknowledgement of the potential impact of missing data elements. A conscious decision was made to either impute the data value from other elements present, to the median, or to the value indicating absence of the risk factor for some of the data elements, and to exclude others. The range of missing value was between 1% and 3.5%. The conclusion was that this did not lead to bias in the measure.

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

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

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

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

The conscious decision regarding social risk factors is discussed in the submission and below in Question 3.

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

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

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

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

c. Any concerns regarding the risk-adjustment approach?
*TIPS: Consider the following: If a justification for **not risk adjusting** is provided, is there any evidence that contradicts the developer's rationale and analysis? If the developer asserts there is **no conceptual basis** for adjusting this measure for social risk factors, do you agree with the rationale? **If risk adjusted**: Are the candidate and final variables included in the risk adjustment model adequately described for the measure to be implemented? Are the candidate and final variables included in the risk adjustment model adequately described for the measure to be implemented? Are all of the risk adjustment variables present at the start of care (if not, do you agree with the rationale)? If social risk factors are not included in the risk-adjustment approach, do you agree with the developer's decision? Is an appropriate risk-adjustment strategy included in the measure (e.g., adequate model discrimination and calibration)? Are all statistical model specifications included, including a "clinical model only" if social risk factors are included in the final model?*
☒Yes (please explain below then go to Question #4)
☐No (go to Question #4)

Social risk factors are not collected in the database and therefore not included in the risk adjustment. It is possible that the data elements collected override other social risk factors, or account for them, but it would be nice to see some statement to that effect. Payer status as a proxy is a part of the database but analysis has not been performed as to its additive value to the model.

4. Analysis of potential threats to validity: Any concerns regarding ability to identify meaningful differences in performance or overall poor performance?
☐Yes (please explain below then go to Question #5)
☒No (go to Question #5)

Despite the above considerations, the large sample size and the historical usage of this database does lead to confidence in the assumptions. And, despite the above, there were statistically meaningful differences in performance demonstrated in the measure score.

5. Analysis of potential threats to validity: Any concerns regarding comparability of results if multiple data sources or methods are specified?
☐Yes (please explain below then go to Question #6)
☐No (go to Question #6)
☒Not applicable (go to Question #6)

Sole data source is the abstracted STS.

6. Analysis of potential threats to validity: Any concerns regarding missing data?
☐Yes (please explain below then go to Question #7)
☒No (go to Question #7)

See Question 1 above and note Section 2b6 describes the analysis and subsequent attribution of missing data elements and efforts to minimize their impact.

Assessment of Measure Testing

7. Was empirical validity testing conducted using the measure as specified and appropriate statistical test?
   Answer no if: face validity; only refer to clinical evidence; only descriptive statistics; only describe process for data management/cleaning/computer programming; testing does not match measure specifications (i.e. data, eMeasure, level, setting, patients).
   ☒Yes (go to Question #10) [NOTE: If appropriate empirical testing has been conducted, then evaluation of face validity is not necessary. Go to Question #8 **only if** there is insufficient information provided to evaluate data element and score-level testing.]
   ☐No (please explain below then go to Question #8)

   Section 2d1.x describes the methodology used to assess the weighting and the effect on the metric of star ratings. It is empirical in that it is applied to hospitals (participants?) with more than 30 lobectomies and results in valid separation between one star and three start ratings for both operative mortality and complication rates. Morbidity is noted to explain more of the variation in the score. Sections 2d2.1 and 2d2.2 describe the derivation of the weight distribution. This fit expert panel expectations.

8. Was face validity systematically assessed by recognized experts to determine agreement on whether the computed performance measure score from the measure as specified can be used to distinguish good and poor quality?
   TIPS: Answer no if: focused on data element accuracy/availability/feasibility/other topics; the degree of consensus and any areas of disagreement not provided/discussed.
   ☒Yes (go to Question #9)
   ☐No (please explain below and rate Question #17: OVERALL VALIDITY as INSUFFICIENT)

   Although not required as the assessment of Question 7 was “YES”, the measure score was assessed by a panel of experts and the methodology was felt to accurately portray the relative contribution of mortality and morbidities to the overall score. It did result in statistically significant differences between those with one-star and three-star ratings.

9. **RATING (face validity)** - Do the face validity testing results indicate substantial agreement that the performance measure score from the measure as specified can be used to distinguish quality AND potential threats to validity are not a problem, OR are adequately addressed so results are not biased?
   ☒Yes (if a NEW measure, rate Question #17: OVERALL VALIDITY as MODERATE)
   ☐ Yes (if a MAINTENANCE measure, do you agree with the justification for not conducting empirical testing? If no, rate Question #17: OVERALL VALIDITY as INSUFFICIENT; otherwise, rate Question #17: OVERALL VALIDITY as MODERATE)
   ☐No (please explain below and rate Question #17: OVERALL VALIDITY AS LOW)

10. Was validity testing conducted with computed performance measure scores for each measured entity?
    TIPS: Answer no if: one overall score for all patients in sample used for testing patient-level data.
    ☒Yes (go to Question #11)
    ☐No (please explain below and go to Question #13)

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

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

☒Yes (go to Question #12)

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

Section 2d1.2 demonstrates that the score results do separate the groups in to high, medium and low performers (star ratings), and that the score does reflect both components of mortality and major complications. The score does demonstrate that the components included in the composite are consistent with the described quality construct and add value to the overall composite.

The question is, when the components of the composite score are THE two most important quality outcomes pertinent to the patient, in this case, mortality and complications, can those themselves be used as indicators of quality from validity testing perspective, which is shown by the table presented. I cannot think of more important quality indicators against which these two could be tested and, therefore, have to say yes to the methodological appropriateness.

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

☐High (go to Question #14)

☒Moderate (go to Question #14)

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

☐Insufficient

It would have been nice to see some data about the effect of different weightings on the validity of the composite score. The predominant test was face validity and the score (and star ratings) derived from the model.

13. Was other validity testing reported?

☒Yes (go to Question #14)

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

Face validity with a panel of experts was used to assess the validity of the model, who said that an 82.7/17.3 ratio was intuitively supported.

14. Was validity testing conducted with <u>patient-level data elements</u>?

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

☒Yes (go to Question #15)

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

Apparently, though not stated, probably due to resource requirements, the data field validity was capped at 500 maximum denominator. It would be nice to state that.

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

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

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

☐Yes (go to Question #16)

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

It is noted that due to the absence of access to all of the data results, a kappa statistic could not be provided. Generally, percent agreement is not sufficient while easily understood. Another option would have been sensitivity/specificity calculations.

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

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

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

☒Insufficient (go to Question #17)

17. OVERALL VALIDITY RATING

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

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

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

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

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

The statistic for data element validity is not the best available. The authors did mention their lack of ability to calculate a kappa statistic. There is a conscious lack of the use of social risk factors. See Question 11 above for the discussion regarding score level validity.

FOR COMPOSITE MEASURES ONLY: Empirical analyses to support composite construction

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

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

☐High

☒Moderate

☐Low (please explain below)

☐Insufficient (please explain below)

See Questions 11, 12 and 13 above. Clinical rationale is good. Precisely how the 83/17 ratio was derived and why is it is the most applicable one is not clear from the data submitted.

Evaluation B

Scientific Acceptability

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

Instructions:

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

Measure Number: 3294
Measure Title: STS Lobectomy for Lung Cancer Composite Score

RELIABILITY

11. Are submitted specifications precise, unambiguous, and complete so that they can be consistently implemented? *NOTE: NQF staff will conduct a separate, more technical, check of eMeasure (eCQM) specifications, value sets, logic, and feasibility, so no need to consider these in your evaluation.*
    TIPS: Consider the following: Are all the data elements clearly defined? Are all appropriate codes included? Is the logic or calculation algorithm clear? Is it likely this measure can be consistently implemented?
    ☒Yes (go to Question #2)
    ☐No (please explain below, and go to Question #2) NOTE that even though ***non-precise***

specifications should result in an overall LOW rating for reliability, we still want you to look at the testing results.

12. Was empirical reliability testing (at the data element or measure score level) conducted using statistical tests with the measure as specified?
    TIPS: Check the 2nd "NO" box below if: only descriptive statistics provided; only describes process for data management/cleaning/computer programming; testing does not match measure specifications (i.e. data, eMeasure, level of analysis, patients)
    ☒Yes (go to Question #4)
    ☐No, there is reliability testing information, but *not* using statistical tests and/or not for the measure as specified OR there is no reliability testing (please explain below then go to Question #3)

13. Was **empirical VALIDITY testing** of patient-level data conducted?
    ☐Yes (use your rating from data element validity testing – Question #16- under Validity Section)
    ☐No (please explain below and rate Question #11: OVERALL RELIABILITY as INSUFFICIENT and proceed to the VALIDITY SECTION)

14. Was reliability testing conducted with computed performance measure scores for each measured entity?
    TIPS: Answer no if: only one overall score for all patients in sample used for testing patient-level data
    ☒Yes (go to Question #5)
    ☐No (go to Question #8)

15. Was the method described and appropriate for assessing the proportion of variability due to real differences among measured entities? *NOTE: If multiple methods used, at least one must be appropriate.*
    TIPS: Examples of appropriate methods include signal-to-noise analysis (e.g. Adams/RAND tutorial); random split-half correlation; other accepted method with description of how it assesses reliability of the performance score.
    ☒Yes (go to Question #6)
    ☐No (please explain below then go to Question #8)

16. **RATING (score level)** - What is the level of certainty or confidence that the performance measure scores are reliable?
    TIPS: Consider the following: Is the test sample adequate to generalize for widespread implementation? Do the results demonstrate sufficient reliability so that differences in performance can be identified?
    ☐High (go to Question #8)
    ☒Moderate (go to Question #8)
    ☐Low (please explain below then go to Question #7)

17. Was other reliability testing reported?
    ☐Yes (go to Question #8)
    ☐No (rate Question #11: OVERALL RELIABILITY as LOW and proceed to the VALIDITY SECTION)

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

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

☐Yes (go to Question #9)

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

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

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

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

☐Yes (go to Question #10)

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

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

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

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

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

☐Insufficient (go to Question #11)

11. OVERALL RELIABILITY RATING

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

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

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

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

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

VALIDITY

Assessment of Threats to Validity

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

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

☒Yes (go to Question #2)

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

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

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

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

☒No (go to Question #3)

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

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

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

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

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

c. Any concerns regarding the risk-adjustment approach?

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

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

☐No (go to Question #4)

Only concern is with the use of a random effects model for a procedure in which there may be a significant volume effect. Because such models can artificially shrink low-volume providers to the mean, they can alter the ordering of performance significantly, and mask any poor performance associated with low volume status.

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

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

☐No (go to Question #5)

As above

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

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

☐No (go to Question #6)

☒Not applicable (go to Question #6)

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

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

☒No (go to Question #7)

Assessment of Measure Testing

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

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

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

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

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

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

☐Yes (go to Question #9)

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

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

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

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

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

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

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

☐Yes (go to Question #11)

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

Confidence interval testing was shown, but there is no validity testing of the measure score that meets the NQF recommendations for such ("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.")

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

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

☐Yes (go to Question #12)

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

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

☐High (go to Question #14)
☐Moderate (go to Question #14)
☐Low (please explain below then go to Question #13)
☐Insufficient

29. Was other validity testing reported?

☒Yes (go to Question #14)
☐No (please explain below and rate Question #17: OVERALL VALIDITY as LOW)

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

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

☒Yes (go to Question #15)
☐No (please explain below and rate Question #17: OVERALL VALIDITY as INSUFFICIENT if no score-level testing was conducted, otherwise, rate Question #17: OVERALL VALIDITY based on score-level rating from Question #12)

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

TIPS: For example: Data validity/accuracy as compared to authoritative source- sensitivity, specificity, PPV, NPV; other accepted method with description of how it assesses validity of the data elements.
Answer No if: only assessed percent agreement; did not assess separately for all data elements (at least numerator, denominator, exclusions)

☐Yes (go to Question #16)
☒No (please explain below and rate Question #16 as INSUFFICIENT)

Only assessed percent agreement –this is OK this time given the high agreement, but will need to use other listed methods above in future submissions.

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

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

Please see note above – ordinarily testing only percent agreement would be unacceptable, but will rate as moderate if measure developers submit more appropriate testing with future submissions.

17. OVERALL VALIDITY RATING

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

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

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

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

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

FOR COMPOSITE MEASURES ONLY: Empirical analyses to support composite construction

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

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

☐High

☒Moderate

☐Low (please explain below)

☐Insufficient (please explain below)

Evaluation C

Scientific Acceptability

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

Instructions:

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

Measure Number: 3294
Measure Title: Risk-Adjusted Morbidity and Mortality for Lung Resection for Lung Cancer

RELIABILITY

21. Are submitted specifications precise, unambiguous, and complete so that they can be consistently implemented? *NOTE: NQF staff will conduct a separate, more technical, check of eMeasure (eCQM) specifications, value sets, logic, and feasibility, so no need to consider these in your evaluation.*
    TIPS: Consider the following: Are all the data elements clearly defined? Are all appropriate codes included? Is the logic or calculation algorithm clear? Is it likely this measure can be consistently implemented?
    ☒Yes (go to Question #2)
    Reliability of data elements was supported by external audit of the General Thoracic Surgery Database (GTSD) demonstrating high agreement rates and validation of data accuracy.
    ☐No (please explain below, and go to Question #2) NOTE that even though ***non-precise specifications should result in an overall LOW rating for reliability***, we still want you to look at the testing results.

22. Was empirical reliability testing (at the data element or measure score level) conducted using statistical tests with the measure as specified?
    TIPS: Check the 2nd "NO" box below if: only descriptive statistics provided; only describes process for data management/cleaning/computer programming; testing does not match measure specifications (i.e. data, eMeasure, level of analysis, patients)
    ☒Yes (go to Question #4)
    ☐No, there is reliability testing information, but *not* using statistical tests and/or not for the measure as specified OR there is no reliability testing (please explain below then go to Question #3)

23. Was **empirical VALIDITY testing** of patient-level data conducted?
    ☐Yes (use your rating from data element validity testing – Question #16- under Validity Section)
    ☐No (please explain below and rate Question #11: OVERALL RELIABILITY as INSUFFICIENT and proceed to the VALIDITY SECTION)

24. Was reliability testing conducted with computed performance measure scores for each measured entity?
    TIPS: Answer no if: only one overall score for all patients in sample used for testing patient-level data
    ☐Yes (go to Question #5)
    ☒No (go to Question #8)

25. Was the method described and appropriate for assessing the proportion of variability due to real differences among measured entities? *NOTE: If multiple methods used, at least one must be appropriate.*
    TIPS: Examples of appropriate methods include signal-to-noise analysis (e.g. Adams/RAND tutorial); random split-half correlation; other accepted method with description of how it assesses reliability of the performance score.
    ☐Yes (go to Question #6)
    ☐No (please explain below then go to Question #8)

26. **RATING (score level)** - What is the level of certainty or confidence that the performance measure scores are reliable?
    TIPS: Consider the following: Is the test sample adequate to generalize for widespread implementation?

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

☐High (go to Question #8)

☐Moderate (go to Question #8)

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

27. Was other reliability testing reported?

☐Yes (go to Question #8)

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

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

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

☒Yes (go to Question #9)

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

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

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

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

☐Yes (go to Question #10)

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

Only agreement rates were provided in the analysis.

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

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

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

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

☒Insufficient (go to Question #11)

11. OVERALL RELIABILITY RATING

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

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

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

☒Low (please explain below) [NOTE: Should rate LOW if you believe specifications are NOT precise,

unambiguous, and complete]

The submitters demonstrated a robust analysis of inter-abstractor agreement across the hospitals examined. Analysis would be much stronger if they obtained the case level data to compute a Kappa statistic to test interrater reliability.

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

VALIDITY

Assessment of Threats to Validity

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

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

☒Yes (go to Question #2)

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

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

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

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

☒No (go to Question #3)

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

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

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

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

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

c. Any concerns regarding the risk-adjustment approach?

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

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

☒No (go to Question #4)

Risk adjustment for the clinical indicators is strongly supported. I agree with the submitters that social risk data is not available in the GTSD but would encourage the Society of Thoracic Surgeons to consider adding social risk factors to their data collection tools. Currently the GTSD does collect Primary and Secondary Payor information which could be used for Dual Eligibility stratification and possibly used as a risk adjustment.

The multivariable logistic models demonstrated statistical significance in all patient level data except Diabetes and Hypertension in all 3 models. I would question the value of leaving these in the models.

36. Analysis of potential threats to validity: Any concerns regarding ability to identify meaningful differences in performance or overall poor performance?
    ☐Yes (please explain below then go to Question #5)
    ☒No (go to Question #5)
    The submitters validated a difference in performance using the Bayesian modeling to compare the Standardized Incidence Ratio between 231 hospitals.

37. Analysis of potential threats to validity: Any concerns regarding comparability of results if multiple data sources or methods are specified?
    ☐Yes (please explain below then go to Question #6)
    ☐No (go to Question #6)
    ☒Not applicable (go to Question #6)

38. Analysis of potential threats to validity: Any concerns regarding missing data?
    ☐Yes (please explain below then go to Question #7)
    ☒No (go to Question #7)
    Investigators adequately address missing data in their analysis.

Assessment of Measure Testing

39. Was empirical validity testing conducted using the measure as specified and appropriate statistical test?
    Answer no if: face validity; only refer to clinical evidence; only descriptive statistics; only describe process for data management/cleaning/computer programming; testing does not match measure specifications (i.e. data, eMeasure, level, setting, patients).
    ☒Yes (go to Question #10) [NOTE: If appropriate empirical testing has been conducted, then evaluation of face validity is not necessary. Go to Question #8 **only if** there is insufficient information provided to evaluate data element and score-level testing.]
    ☐No (please explain below then go to Question #8)

40. Was face validity systematically assessed by recognized experts to determine agreement on whether the computed performance measure score from the measure as specified can be used to distinguish good and poor quality?
    TIPS: Answer no if: focused on data element accuracy/availability/feasibility/other topics; the degree of consensus and any areas of disagreement not provided/discussed.
    ☐Yes (go to Question #9)
    ☐No (please explain below and rate Question #17: OVERALL VALIDITY as INSUFFICIENT)

41. **RATING (face validity)** - Do the face validity testing results indicate substantial agreement that the performance measure score from the measure as specified can be used to distinguish quality AND potential threats to validity are not a problem, OR are adequately addressed so results are not biased?
    ☐Yes (if a NEW measure, rate Question #17: OVERALL VALIDITY as MODERATE)
    ☐ Yes (if a MAINTENANCE measure, do you agree with the justification for not

conducting empirical testing? If no, rate Question #17: OVERALL VALIDITY as INSUFFICIENT; otherwise, rate Question #17: OVERALL VALIDITY as MODERATE)

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

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

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

☐Yes (go to Question #11)

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

No evidence that validity of performance score was tested. If the submitters have performed performance score testing for their previous risk-adjusted models, I would recommend updating the performance score testing with the proposed risk adjusted models.

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

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

☐Yes (go to Question #12)

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

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

☐High (go to Question #14)

☐Moderate (go to Question #14)

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

☐Insufficient

45. Was other validity testing reported?

☒Yes (go to Question #14)

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

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

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

☒Yes (go to Question #15)

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

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

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

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

☐Yes (go to Question #16)

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

Only agreement rates were assessed.

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

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

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

It would be a much stronger analysis if the developer obtained the case level results to provide a kappa statistic.

☐Insufficient (go to Question #17)

17. OVERALL VALIDITY RATING

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

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

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

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

No testing for threats to validity evident in the information provided by the submitters

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

FOR COMPOSITE MEASURES ONLY: Empirical analyses to support composite construction

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

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

☐High

☒Moderate

The statistical analysis supports the use of the Mortality or Major Morbidity Composite Model for risk adjustment and performance measurement. Although, the referenced article did show only fair performance of the composite model using the C-statistic results. I would recommend the submitters include the referenced article in their submission materials.

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

☐Low (please explain below)

☐Insufficient (please explain below)

Evaluation D

Scientific Acceptability

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

Instructions:

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

Measure Number: 3294
Measure Title: STS Lobectomy for Lung Cancer Composite Score

RELIABILITY

31. Are submitted specifications precise, unambiguous, and complete so that they can be consistently implemented? *NOTE: NQF staff will conduct a separate, more technical, check of eMeasure (eCQM) specifications, value sets, logic, and feasibility, so no need to consider these in your evaluation.*
    TIPS: Consider the following: Are all the data elements clearly defined? Are all appropriate codes included? Is the logic or calculation algorithm clear? Is it likely this measure can be consistently implemented?
    ☒Yes (go to Question #2)
    ☐No (please explain below, and go to Question #2) NOTE that even though ***non-precise specifications should result in an overall LOW rating for reliability***, we still want you to look at the testing results.

32. Was empirical reliability testing (at the data element or measure score level) conducted using statistical tests with the measure as specified?
    TIPS: Check the 2nd "NO" box below if: only descriptive statistics provided; only describes process for data management/cleaning/computer programming; testing does not match measure specifications (i.e. data, eMeasure, level of analysis, patients)
    ☒Yes (go to Question #4)
    ☐No, there is reliability testing information, but *not* using statistical tests and/or not for the measure as specified OR there is no reliability testing (please explain below then go to Question #3)

33. Was **empirical VALIDITY testing** of patient-level data conducted?
    ☐Yes (use your rating from data element validity testing – Question #16- under Validity Section)
    ☐No (please explain below and rate Question #11: OVERALL RELIABILITY as INSUFFICIENT and proceed to the VALIDITY SECTION)

34. Was reliability testing conducted with computed performance measure scores for each measured entity?
    TIPS: Answer no if: only one overall score for all patients in sample used for testing patient-level data
    ☒Yes (go to Question #5)
    ☐No (go to Question #8)

35. Was the method described and appropriate for assessing the proportion of variability due to real differences among measured entities? *NOTE: If multiple methods used, at least one must be appropriate.*
    TIPS: Examples of appropriate methods include signal-to-noise analysis (e.g. Adams/RAND tutorial); random split-half correlation; other accepted method with description of how it assesses reliability of the performance score.
    ☒Yes (go to Question #6)
    ☐No (please explain below then go to Question #8)

36. **RATING (score level)** - What is the level of certainty or confidence that the performance measure scores are reliable?
    TIPS: Consider the following: Is the test sample adequate to generalize for widespread implementation? Do the results demonstrate sufficient reliability so that differences in performance can be identified?

☐High (go to Question #8)
☒Moderate (go to Question #8)
☐Low (please explain below then go to Question #7)

37. Was other reliability testing reported?
☐Yes (go to Question #8)
☐No (rate Question #11: OVERALL RELIABILITY as LOW and proceed to the VALIDITY SECTION)

38. Was reliability testing conducted with patient-level data elements that are used to construct the performance measure?
TIPS: Prior reliability studies of the same data elements may be submitted; if comparing abstraction to "authoritative source/gold standard" see Validity Section Question #15)
☐Yes (go to Question #9)
☒No (if there is score-level testing, rate Question #11: OVERALL RELIABILITY based on score-level rating from Question #6; otherwise, rate Question #11: OVERALL RELIABILITY as INSUFFICIENT. Then proceed to the VALIDITY SECTION)

39. Was the method described and appropriate for assessing the reliability of ALL critical data elements?
TIPS: For example: inter-abstractor agreement (ICC, Kappa); other accepted method with description of how it assesses reliability of the data elements
Answer no if: only assessed percent agreement; did not assess separately for all data elements (at least numerator, denominator, exclusions)
☐Yes (go to Question #10)
☐No (if no, please explain below and rate Question #10 as INSUFFICIENT)

My one concern with the reliability of the data elements is changes in the registry reporting platform over time. Opening up new reporting options may reduce reliability of data over time.

40. **RATING (data element)** – Based on the reliability statistic and scope of testing (number and representativeness of patients and entities), what is the level of certainty or confidence that the data used in the measure are reliable?
TIPS: Consider the following: Is the test sample adequate to generalize for widespread implementation? Can data elements be collected consistently?
☐Moderate (if score-level testing was NOT conducted, rate Question #11: OVERALL RELIABILITY as MODERATE)
☐Low (if score-level testing was NOT conducted, rate Question #11: OVERALL RELIABILITY as LOW)
☐Insufficient (go to Question #11)

11. OVERALL RELIABILITY RATING

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

☐High (NOTE: Can be HIGH only if score-level testing has been conducted)
☒Moderate (NOTE: Moderate is the highest eligible rating if score-level testing has not been conducted)
☐Low (please explain below) [NOTE: Should rate LOW if you believe specifications are NOT precise,

unambiguous, and complete]

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

The measure is tested at the hospital level. The measure summary form indicates that it can be used for hospitals or group practices, but I do not see any evidence of reliability testing with group practice data.

VALIDITY

Assessment of Threats to Validity

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

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

☒Yes (go to Question #2)

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

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

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

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

☐No (go to Question #3)

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

I was slightly concerned about dropping cases with missing discharge mortality status because I cannot tell if this introduces selection bias or offers an opportunity for gaming. I'm assuming this is a relatively rare event, although I didn't see the number of cases dropped in either the Composite Measure Testing worksheet or the journal article.

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

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

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

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

c. Any concerns regarding the risk-adjustment approach?

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

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

☒No (go to Question #4)

The one thing to note with a Bayesian risk adjustment model is the tendency for scores to fall in the middle of the distribution. We see this here with 91.4 percent of cases ending up with 2 stars. One strength of the risk model is that covariates were selected on an a-priori or theoretical basis and retained in the model regardless of impact rather than through a data driven process. The model c-statistics are modest, but not unexpected for clinical data.

52. Analysis of potential threats to validity: Any concerns regarding ability to identify meaningful differences in performance or overall poor performance?
    ☐Yes (please explain below then go to Question #5)
    ☒No (go to Question #5)

53. Analysis of potential threats to validity: Any concerns regarding comparability of results if multiple data sources or methods are specified?
    ☐Yes (please explain below then go to Question #6)
    ☐No (go to Question #6)
    ☒Not applicable (go to Question #6)

54. Analysis of potential threats to validity: Any concerns regarding missing data?
    ☒Yes (please explain below then go to Question #7)
    ☐No (go to Question #7)

As I mentioned above, I have concerns about potential selection bias for sites with missing mortality information. It would be helpful to know the number of excluded cases – I assume it is small and random.

Assessment of Measure Testing

55. Was empirical validity testing conducted using the measure as specified and appropriate statistical test?
    Answer no if: face validity; only refer to clinical evidence; only descriptive statistics; only describe process for data management/cleaning/computer programming; testing does not match measure specifications (i.e. data, eMeasure, level, setting, patients).
    ☒Yes (go to Question #10) [NOTE: If appropriate empirical testing has been conducted, then evaluation of face validity is not necessary. Go to Question #8 **only if** there is insufficient information provided to evaluate data element and score-level testing.]
    ☐No (please explain below then go to Question #8)

56. Was face validity systematically assessed by recognized experts to determine agreement on whether the computed performance measure score from the measure as specified can be used to distinguish good and poor quality?
    TIPS: Answer no if: focused on data element accuracy/availability/feasibility/other topics; the degree of consensus and any areas of disagreement not provided/discussed.
    ☐Yes (go to Question #9)
    ☐No (please explain below and rate Question #17: OVERALL VALIDITY as INSUFFICIENT)

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

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

58. Was validity testing conducted with computed performance measure scores for each measured entity?
TIPS: Answer no if: one overall score for all patients in sample used for testing patient-level data.
☒Yes (go to Question #11)
☐No (please explain below and go to Question #13)

59. Was the method described and appropriate for assessing conceptually and theoretically sound hypothesized relationships?
TIPS: For example: correlation of the performance measure score on this measure and other performance measures; differences in performance scores between groups known to differ on quality; other accepted method with description of how it assesses validity of the performance score
☒Yes (go to Question #12)
☐No (please explain below, rate Question #12 as INSUFFICIENT and then go to Question #14)

This was done by looking at the relationship between observed rates of the two outcomes (mortality and major complications) and the overall star rating for the hospital. As the authors point out, there is a clear linear relationship between observed components. Worth noting, the 95% confidence intervals for the mortality measure almost overlap for the 1-star and 2-star groups. If hospitals with lower volume were included in the analysis these two groups may not be distinct.

Grouping measure scores by star rating helps confirm that the composite is not driven by a single measure and that both measures move together. However, as the authors point out, the major morbidity measure drives the variance. This is not surprising since it is made up of 9 medical complications and is itself a composite of sorts. It would be helpful to see a confirmatory factor analysis or structural measurement model to better understand how all 10 items relate to each other. Finally, it would be beneficial to have an external measure of adverse events after lobectomy or a broader category of lung surgeries to group hospitals (i.e., an independent measure that is not part of the composite).

60. **RATING (measure score)** - Based on the measure score results (significance, strength) and scope of testing (number of measured entities and representativeness) and analysis of potential threats, what is the level of certainty or confidence that the performance measure scores are a valid indicator of quality?
☐High (go to Question #14)
☒Moderate (go to Question #14)
☐Low (please explain below then go to Question #13)
☐Insufficient

61. Was other validity testing reported?
☐Yes (go to Question #14)
☐No (please explain below and rate Question #17: OVERALL VALIDITY as LOW)

62. Was validity testing conducted with patient-level data elements?
TIPS: Prior validity studies of the same data elements may be submitted
☒Yes (go to Question #15)

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

The authors provide information on overall validity testing for the General Thoracic Surgery Database in 2016, 2011 and 2010. In the narrative they refer to auditing 10% of sites for completeness, but only 15 lobectomy cases for accuracy. This leads to confusion with the table shown on pages 9-10 that shows a total of 500 cases for many data elements. It is not clear what this table is reporting at the data element level.

63. Was the method described and appropriate for assessing the accuracy of ALL critical data elements? *NOTE that data element validation from the literature is acceptable.*
TIPS: For example: Data validity/accuracy as compared to authoritative source- sensitivity, specificity, PPV, NPV; other accepted method with description of how it assesses validity of the data elements.
Answer No if: only assessed percent agreement; did not assess separately for all data elements (at least numerator, denominator, exclusions)
☒Yes (go to Question #16)
☐No (please explain below and rate Question #16 as INSUFFICIENT)

The method is appropriate, but as noted above, it is difficult to know if the agreement rates shown in the table are correct given the miss-match between the numbers in the table and text. It is also not clear why there is no one who has both the auditor's rating and the site level data to calculate a kappa statistic. This seems like a key component in assessing and maintain database integrity over time.

64. **RATING (data element)** - Based on the data element testing results (significance, strength) and scope of testing (number and representativeness of patients and entities) and analysis of potential threats, what is the level of certainty or confidence that the data used in the measure are valid?
☐Moderate (if score-level testing was NOT conducted, rate Question #17: OVERALL VALIDITY as MODERATE)
☒Low (please explain below) (if score-level testing was NOT conducted, rate Question #17: OVERALL VALIDITY as LOW)
☐Insufficient (go to Question #17)

This rating is based on the concerns with the miss-match between the agreement rates in the text and tables as well as the lack of a kappa statistic. In addition, it would be helpful to know if the current data reporting options and auditing requirements for 2016 will carry forward to 2017 and beyond. Changes in these methods could adversely affect the validity of future data in the STS database.

17. OVERALL VALIDITY RATING

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

☐High (NOTE: Can be HIGH only if score-level testing has been conducted)
☒Moderate (NOTE: Moderate is the highest eligible rating if score-level testing has NOT been conducted)
☐Low (please explain below) [NOTE: Should rate LOW if you believe that there are threats to validity and/or threats to validity were not assessed]
☐Insufficient (if insufficient, please explain below) [NOTE: For most measure types, testing at both the score level and the data element level is not required] [NOTE: If rating is INSUFFICIENT for all empirical testing, then go back to Question #8 and evaluate any face validity that was conducted, then reconsider this overall rating.]

FOR COMPOSITE MEASURES ONLY: Empirical analyses to support composite construction

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

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

☐High

☒Moderate

☐Low (please explain below)

☐Insufficient (please explain below)

Overall this is a well thought out measure. It builds on the STS registry, which captures the vast majority of cases among participating members and is subjected to an independent auditing process. As the authors point out, not all lobectomies are performed by cardio-thoracic surgeons. From a 'public benefit' perspective, it would be helpful to include all relevant surgeries in the measure, not just the ones performed by a specific type of surgeon. Obviously this is not possible with the risk adjustment model used for the measure, but would be something to consider for the future.

It is important that the measure includes a minimum number of cases (N=30) since the reliability is modest for low case volumes. The composite score is a logical combination of a number of closely related outcomes. The standardization and weighting are strengths of the overall measure. The reliability testing was appropriate and shows modest reliability with relatively low sample sizes. The distribution of participant's composite scores for lobectomy in Figure 1 of Kozower et al. (2016) shows graphically that the measure is able to differentiate performance above and below the mean. Worth noting, composite scores are already relatively high, offering relatively limited room for improvement. Also the Bayesian risk adjustment results push many hospitals to the middle of the distribution, resulting in clear differentiation between high and low performers for a relatively small percent of the overall sample. Finally, this is a composite measure made up of two different measures, each of which captures adverse events after surgery. The measure could be improved by allowing all 10 adverse events to be standardized and weighted individually.

Kozower BD, O'Brein SM, Kosinski AS, et al. (2016). The Society of Thoracic Surgeons Composite Score for Rating Program Performance for Lobectomy for Lung Cancer. Annals of Thoracic Surgery, 101: 1379-1387.

Evaluation E

Scientific Acceptability

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

Instructions:

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

Measure Number: **3294**
Measure Title: STS Lobectomy for Lung Cancer Composite Score

RELIABILITY

41. Are submitted specifications precise, unambiguous, and complete so that they can be consistently implemented? *NOTE: NQF staff will conduct a separate, more technical, check of eMeasure (eCQM) specifications, value sets, logic, and feasibility, so no need to consider these in your evaluation.*
    TIPS: Consider the following: Are all the data elements clearly defined? Are all appropriate codes included? Is the logic or calculation algorithm clear? Is it likely this measure can be consistently implemented?
    ☒Yes (go to Question #2)
    ☐No (please explain below, and go to Question #2) NOTE that even though ***non-precise specifications should result in an overall LOW rating for reliability***, we still want you to look at the testing results.

42. Was empirical reliability testing (at the data element or measure score level) conducted using statistical tests with the measure as specified?
    TIPS: Check the 2nd "NO" box below if: only descriptive statistics provided; only describes process for data management/cleaning/computer programming; testing does not match measure specifications (i.e. data, eMeasure, level of analysis, patients)
    ☒Yes (go to Question #4)
    ☐No, there is reliability testing information, but *not* using statistical tests and/or not for the measure as specified OR there is no reliability testing (please explain below then go to Question #3)

43. Was **empirical VALIDITY testing** of patient-level data conducted?
    ☐Yes (use your rating from data element validity testing – Question #16- under Validity Section)
    ☐No (please explain below and rate Question #11: OVERALL RELIABILITY as INSUFFICIENT and proceed to the VALIDITY SECTION)

44. Was reliability testing conducted with computed performance measure scores for each measured entity?
    TIPS: Answer no if: only one overall score for all patients in sample used for testing patient-level data
    ☒Yes (go to Question #5)
    ☐No (go to Question #8)

45. Was the method described and appropriate for assessing the proportion of variability due to real differences among measured entities? *NOTE: If multiple methods used, at least one must be appropriate.*
    TIPS: Examples of appropriate methods include signal-to-noise analysis (e.g. Adams/RAND tutorial); random split-half correlation; other accepted method with description of how it assesses reliability of the performance score.
    ☒Yes (go to Question #6)
    ☐No (please explain below then go to Question #8)

46. **RATING (score level)** - What is the level of certainty or confidence that the performance measure scores are reliable?
    TIPS: Consider the following: Is the test sample adequate to generalize for widespread implementation? Do the results demonstrate sufficient reliability so that differences in performance can be identified?
    ☐High (go to Question #8)
    ☒Moderate (go to Question #8)
    ☐Low (please explain below then go to Question #7)

47. Was other reliability testing reported?

☐Yes (go to Question #8)

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

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

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

☐Yes (go to Question #9)

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

Data not provided in the submission, but may be available in STS database.

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

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

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

☐Yes (go to Question #10)

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

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

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

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

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

☐Insufficient (go to Question #11)

11. OVERALL RELIABILITY RATING

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

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

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

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

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

It appears reliability is moderately good, and improves, as expected, with increasing number of cases.

VALIDITY

Assessment of Threats to Validity

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

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

☒Yes (go to Question #2)

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

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

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

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

☒No (go to Question #3)

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

No major concerns, but social disparities explicitly ignored, with explanation.

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

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

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

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

c. Any concerns regarding the risk-adjustment approach?

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

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

☒No (go to Question #4)

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

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

☒No (go to Question #5)

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

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

☐No (go to Question #6)
☒Not applicable (go to Question #6)

70. Analysis of potential threats to validity: Any concerns regarding missing data?
☐Yes (please explain below then go to Question #7)
☒No (go to Question #7)

Assessment of Measure Testing

71. Was empirical validity testing conducted using the measure as specified and appropriate statistical test?
Answer no if: face validity; only refer to clinical evidence; only descriptive statistics; only describe process for data management/cleaning/computer programming; testing does not match measure specifications (i.e. data, eMeasure, level, setting, patients).
☒Yes (go to Question #10) [NOTE: If appropriate empirical testing has been conducted, then evaluation of face validity is not necessary. Go to Question #8 **only if** there is insufficient information provided to evaluate data element and score-level testing.]
☐No (please explain below then go to Question #8)

72. Was face validity systematically assessed by recognized experts to determine agreement on whether the computed performance measure score from the measure as specified can be used to distinguish good and poor quality?
TIPS: Answer no if: focused on data element accuracy/availability/feasibility/other topics; the degree of consensus and any areas of disagreement not provided/discussed.
☐Yes (go to Question #9)
☐No (please explain below and rate Question #17: OVERALL VALIDITY as INSUFFICIENT)

73. **RATING (face validity)** - Do the face validity testing results indicate substantial agreement that the performance measure score from the measure as specified can be used to distinguish quality AND potential threats to validity are not a problem, OR are adequately addressed so results are not biased?
☐Yes (if a NEW measure, rate Question #17: OVERALL VALIDITY as MODERATE)
☐ Yes (if a MAINTENANCE measure, do you agree with the justification for not conducting empirical testing? If no, rate Question #17: OVERALL VALIDITY as INSUFFICIENT; otherwise, rate Question #17: OVERALL VALIDITY as MODERATE)
☐No (please explain below and rate Question #17: OVERALL VALIDITY AS LOW)

74. Was validity testing conducted with computed performance measure scores for each measured entity?
TIPS: Answer no if: one overall score for all patients in sample used for testing patient-level data.
☐Yes (go to Question #11)
☒No (please explain below and go to Question #13)
NQF recommends testing hypotheses that the measure scores indicate quality of care, e.g., measure scores differ by groups known to have differences in quality assessed by another valid quality measure or method; or by correlation of measure scores with another valid indicator of quality for a specific topic; or relationship to conceptually similar measures. This submission reported (mostly) separated confidence intervals but no 'anchor' against which to judge validity.

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

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

☐Yes (go to Question #12)

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

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

☐High (go to Question #14)

☐Moderate (go to Question #14)

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

☐Insufficient

77. Was other validity testing reported?

☒Yes (go to Question #14)

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

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

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

☒Yes (go to Question #15)

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

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

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

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

☐Yes (go to Question #16)

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

Kappa statistic for case-level data would help clarify confusion in the submission regarding agreement rates. This can probably be clarified in a follow-up submission

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

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

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

☐Insufficient (go to Question #17)

17. OVERALL VALIDITY RATING

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

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

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

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

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

I suspect that further detail from available information will render this as an acceptable, reliable and valid measure

FOR COMPOSITE MEASURES ONLY: Empirical analyses to support composite construction

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

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

☐High

☒Moderate

☐Low (please explain below)

☐Insufficient (please explain below)

Some concern that there are wide and almost overlapping confidence intervals for the mortality outcome between 1-star and 2-star hospitals. Low volume hospitals, with lower reliability, would likely overlap. With so many hospitals classified in the middle group, this may not be a highly-differentiating outcome measure at the end of the day….but it seems conceptually and structurally sound

NATIONAL QUALITY FORUM—Evidence (subcriterion 1a)

Measure Number (*if previously endorsed*)**:** Click here to enter NQF number
Measure Title: STS Lobectomy for Lung Cancer Composite Score
IF the measure is a component in a composite performance measure, provide the title of the Composite Measure here: Click here to enter composite measure #/ title
Date of Submission: 11/15/2017

Instructions

- *Complete 1a.1 and 1a.2 for all measures. If instrument-based measure, complete 1a.3.*
- *Complete **EITHER 1a.2, 1a.3 or 1a.4** as applicable for the type of measure and evidence.*
- *For composite performance measures:*
  - *A separate evidence form is required for each component measure unless several components were studied together.*
  - *If a component measure is submitted as an individual performance measure, attach the evidence form to the individual measure submission.*
- All information needed to demonstrate meeting the evidence subcriterion (1a) must be in this form. An appendix of *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.

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

1a. Evidence to Support the Measure Focus

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

- Outcome: [3] Empirical data demonstrate a relationship between the outcome and at least one healthcare structure, process, intervention, or service. If not available, wide variation in performance can be used as evidence, assuming the data are from a robust number of providers and results are not subject to systematic bias.
- Intermediate clinical outcome: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured intermediate clinical outcome leads to a desired health outcome.
- Process: [5] a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured process leads to a desired health outcome.
- Structure: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured structure leads to a desired health outcome.

- Efficiency: [6] evidence not required for the resource use component.
- For measures derived from patient reports, evidence should demonstrate that the target population values the measured outcome, process, or structure and finds it meaningful.
- Process measures incorporating Appropriate Use Criteria: See NQF's guidance for evidence for measures, in general; guidance for measures specifically based on clinical practice guidelines apply as well.

Notes

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

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

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

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

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

Outcome

☒ Outcome: Two domains of outcomes are measured: 1. Operative Mortality (death during the same hospitalization as surgery or within 30 days of the procedure), and 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.

☐Patient-reported outcome (PRO): Click here to name the PRO

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

☐ Intermediate clinical outcome (*e.g., lab value*): Click here to name the intermediate outcome

☐ Process: Click here to name what is being measured

☐ Appropriate use measure: Click here to name what is being measured

☐ Structure: Click here to name the structure

☐ Composite: Click here to name what is being measured

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

Postoperative complications and operative mortality are important negative outcomes associated with lung cancer resection surgery, including lobectomy, the most frequently performed lung resection procedure. The STS lung cancer resection risk model (Fernandez et al, 2016) identifies predictors of these outcomes, including patient age, smoking status, comorbid medical conditions, and other patient characteristics, as well as operative approach and the extent of pulmonary resection. Knowledge of these predictors informs clinical decision making by enabling physicians and patients to understand the associations between individual patient characteristics and outcomes and – with continuous feedback of performance data over time – fosters quality improvement.

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

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

n/a

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

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

Data in the STS General Thoracic Surgery Database (GTSD) have 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.

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.

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

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

☐ Clinical Practice Guideline recommendation (with evidence review)

☐ US Preventive Services Task Force Recommendation

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

☐ Other

Source of Systematic Review: • **Title** • **Author** • **Date** • **Citation, including page number** • **URL**	
Quote the guideline or recommendation verbatim about the process, structure or intermediate outcome being measured. If not a guideline, summarize the conclusions from the SR.	

Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	
Provide all other grades and definitions from the evidence grading system	
Grade assigned to the **recommendation** with definition of the grade	
Provide all other grades and definitions from the recommendation grading system	
Body of evidence: • Quantity – how many studies? • Quality – what type of studies?	
Estimates of benefit and consistency across studies	
What harms were identified?	
Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	

1a.4 OTHER SOURCE OF EVIDENCE

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

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

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

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

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

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

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

NQF_evidence_attachment_STS-3294-111517.docx

1a.1 For Maintenance of Endorsement: Is there new evidence about the measure since the last update/submission?

Do not remove any existing information. If there have been any changes to evidence, the Committee will consider the new evidence. Please use the most current version of the evidence attachment (v7.1). Please use red font to indicate updated evidence.

No

1b. Performance Gap

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

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

1b.1. Briefly explain the rationale for this measure *(e.g., how the measure will improve the quality of care, the benefits or improvements in quality envisioned by use of this measure)*
If a COMPOSITE (e.g., combination of component measure scores, all-or-none, any-or-none), SKIP this question and answer the composite questions.
n/a

1b.2. Provide performance scores on the measure as specified (current and over time) at the specified level of analysis. *(This is required for maintenance of endorsement. Include mean, std dev, min, max, interquartile range, scores by decile. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities include.) This information also will be used to address the sub-criterion on improvement (4b1) under Usability and Use.*
The measure was calculated in two overlapping 3-year time periods, January 1, 2014 – December 31, 2016 and January 1, 2013 – December 31, 2015. 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 1, 2013 – December 31, 2015		January 1, 2014 – December 31, 2016	
	All participants	=30 cases	All participants	=30 cases
No. of participants	242	185	233	186
No. of operations	23594	22752	24912	24318
Mean	0.972	0.972	0.973	0.974
SD	0.007	0.008	0.006	0.007
IQR	0.008	0.009	0.007	0.009
Minimum	0.945	0.945	0.953	0.953
10%	0.961	0.96	0.965	0.965
20%	0.967	0.967	0.969	0.968
30%	0.97	0.969	0.971	0.971
40%	0.971	0.971	0.973	0.973
50%	0.973	0.973	0.974	0.975
60%	0.974	0.975	0.976	0.976
70%	0.975	0.976	0.977	0.977
80%	0.977	0.978	0.979	0.979
90%	0.979	0.98	0.981	0.982
Maximum	0.988	0.988	0.987	0.987
Midwest	49	38	48	38
Northeast	71	52	67	51
South	82	65	82	67
West	40	30	36	30

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

1b.4. Provide disparities data from the measure as specified (current and over time) by population group, e.g., by race/ethnicity, gender, age, insurance status, socioeconomic status, and/or disability. *(This is required for maintenance of endorsement. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included.) For measures that show high levels of performance, i.e., "topped out", disparities data may demonstrate an opportunity for improvement/gap in care for certain sub-populations. This information also will be used to address the sub-criterion on improvement (4b1) under Usability and Use.*
DATES: Jan. 1, 2014 - Dec. 31, 2016

INCIDENCE N= 33,326

DEMOGRAPHICS
Age (years)
Mean 65.7
Median 67.0
25th Percentile 59.0
75th Percentile 73.0

Gender, Female 54.7%

Race
Caucasian 84.9%
Black 8.8%
Asian 2.9%
Native American 0.3%
Native Hawaiian/Pac Islander 0.2%
Other 2.5%
Multiple Races 0.7%
Missing/unknown/not documented 1.1%

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

1c. Composite Quality Construct and Rationale

1c.1. **A composite performance measure is a combination of two or more component measures, each of which individually reflects 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);

1c.1. Please identify the composite measure construction: two or more individual performance measure scores combined into one score

1c.2. **Describe the quality construct, including:**

- the overall area of quality
- included component measures and
- the relationship of the component measures to the overall composite and to each other.

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

1c.3. **Describe the rationale for constructing a composite measure, including how the composite provides a distinctive or additive value over the component measures individually.**
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.

1c.4. **Describe how the aggregation and weighting of the component measures are consistent with the stated quality construct and rationale.**
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.

NATIONAL QUALITY FORUM—Composite Measure Testing (subcriteria 2a2, 2b1-2b6)

Measure Number (*if previously endorsed*)**:**
Composite Measure Title: STS Lobectomy for Lung Cancer Composite Score
Date of Submission: 11/15/2017
Composite Construction:
☒Two or more individual performance measure scores combined into one score
☐ All-or-none measures (e.g., all essential care processes received or outcomes experienced by each patient)

Instructions: Please contact NQF staff before you begin.

- If a component measure is submitted as an individual performance measure, the non-composite measure testing form must also be completed and attached to the individual measure 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.
- **Sections 1, 2a2, 2b1, 2b2, and 2b4 must be completed.**
- **For composites with outcome and resource use measures**, section **2b3** also must be completed.
- If specified for **multiple data sources/sets of specificaitons** (e.g., claims and EHRs), section **2b5** also must be completed.
- Respond to all questions as instructed with answers immediately following the question. All information on testing to demonstrate meeting the subcriteria for reliability (2a2) and validity (2b1-2b6) and composites (2c) must be in this form. An appendix for *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 25 pages (*incuding questions/instructions;* minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.

- For information on the most updated guidance on how to address social risk factors variables and testing in this form refer to the release notes for version 7.1 of the Measure Testing Attachment. and the 2017 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.

2a2. Reliability testing [10] demonstrates the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise. For **instrument-based measures** (including **PRO-PMs**) **and composite performance measures**, reliability should be demonstrated for the computed performance score.

2b1. Validity testing [11] demonstrates that the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality. For **instrument based measures (including PRO-PMs) and composite performance measures**, validity should be demonstrated for the computed performance score.

2b2. Exclusions are supported by the clinical evidenceand are of sufficient frequency to warrant inclusion in the specifications of the measure; [12]

AND

If patient preference (e.g., informed decisionmaking) is a basis for exclusion, there must be evidence that the exclusion impacts performance on the measure; in such cases, the measure must be specified so that the information about patient preference and the effect on the measure is transparent (e.g., numerator category computed separately, denominator exclusion category computed separately). [13]

2b3. For outcome measures and other measures when indicated (e.g., resource use):

- **an evidence-based risk-adjustment strategy** (e.g., risk models, risk stratification) is specified; is based on patient factors (including clinical and social risk factors) that influence the measured outcome and are present at start of care; [14,15] and has demonstrated adequate discrimination and calibration

OR

- rationale/data support no risk adjustment/ stratification.

2b4. Data analysis of computed measure scores demonstrates that methods for scoring and analysis of the specified measure allow for **identification of statistically significant and practically/clinically meaningful** [16] **differences in performance**;

OR

there is evidence of overall less-than-optimal performance.

2b5. If multiple data sources/methods are specified, there is demonstration they produce comparable results.

2b6. Analyses identify the extent and distribution of **missing data** (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias.

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)*

Notes

10. Reliability testing applies to both the data elements and computed measure score. Examples of reliability testing for data elements include, but are not limited to: inter-rater/abstractor or intra-rater/abstractor studies; internal consistency for multi-item scales; test-retest for survey items. Reliability testing of the measure score addresses precision of measurement (e.g., signal-to-noise).

11. Validity testing applies to both the data elements and computed measure score. Validity testing of data elements typically analyzes agreement with another authoritative source of the same information. Examples of validity testing of the measure score include, but are not limited to: testing hypotheses that the measures scores indicate quality of care, e.g., measure scores are different for groups known to have differences in quality assessed by another valid quality measure or method; correlation of measure scores with another valid indicator of quality for the specific topic; or relationship to conceptually related measures (e.g., scores on process measures to scores on outcome measures). Face validity of the measure score as a quality indicator may be adequate if accomplished through a systematic and transparent process, by identified experts, and explicitly addresses whether performance scores resulting from the measure as specified can be used to distinguish good from poor quality. The degree of consensus and any areas of disagreement must be provided/discussed.

12. Examples of evidence that an exclusion distorts measure results include, but are not limited to: frequency of occurrence, variability of exclusions across providers, and sensitivity analyses with and without the exclusion.

13. Patient preference is not a clinical exception to eligibility and can be influenced by provider interventions.

14. Risk factors that influence outcomes should not be specified as exclusions.

15. With large enough sample sizes, small differences that are statistically significant may or may not be practically or clinically meaningful. The substantive question may be, for example, whether a statistically significant difference of one percentage point in the percentage of patients who received smoking cessation counseling (e.g., 74 percent v. 75 percent) is clinically meaningful; or whether a statistically significant difference of $25 in cost for an episode of care (e.g., $5,000 v. $5,025) is practically meaningful. Measures with overall less-than-optimal performance may not demonstrate much variability across providers.

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

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

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

Measure Specified to Use Data From:	Measure Tested with Data From:

(*must be consistent with data sources entered in S.17*)	
☐ abstracted from paper record	☐ abstracted from paper record
☐ claims	☐ claims
☒ registry	☒ registry
☐ abstracted from electronic health record	☐ abstracted from electronic health record
☐ eMeasure (HQMF) implemented in EHRs	☐ eMeasure (HQMF) implemented in EHRs
☐ other:	☐ other:

1.2. If an existing dataset was used, identify the specific dataset (*the dataset used for testing must be consistent with the measure specifications for target population and healthcare entities being measured; e.g., Medicare Part A claims, Medicaid claims, other commercial insurance, nursing home MDS, home health OASIS, clinical registry*).
STS General Thoracic Surgery Database, Version 2.3

1.3. What are the dates of the data used in testing? 01/01/2014 – 12/31/2016

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

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

1.5. How many and which <u>measured entities</u> were included in the testing and analysis (by level of analysis and data source)? (*identify the number and descriptive characteristics of measured entities included in the analysis (e.g., size, location, type); if a sample was used, describe how entities were selected for inclusion in the sample*)
The analysis population consisted of all STS records for patients meeting measure inclusion criteria who had their surgery during January 1, 2014 through December 31, 2016. The population included 24,912 patient records from 233 hospitals.

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

Includes 24,912 eligible patients. Patient characteristics are below.

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%
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%

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

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

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

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

2a2. RELIABILITY TESTING

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

2a2.1. What level of reliability testing was conducted? (*may be one or both levels*)

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

☒ **Performance measure score** (e.g., *signal-to-noise analysis*)

2a2.2. Describe the method of reliability testing and what it tests (*describe the steps—do not just name a method; what type of error does it test; what statistical analysis was used*)

Reliability is conventionally defined as the proportion of variation in a 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, $\theta_1,\ldots,\theta_N$, that is:

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

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

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

where

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

with $\theta_h^{(l)}$ 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.

2a2.3. What were the statistical results from reliability testing? (e.*g.*, *percent agreement and kappa for the critical data elements; distribution of reliability statistics from a signal-to-noise analysis*)

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.

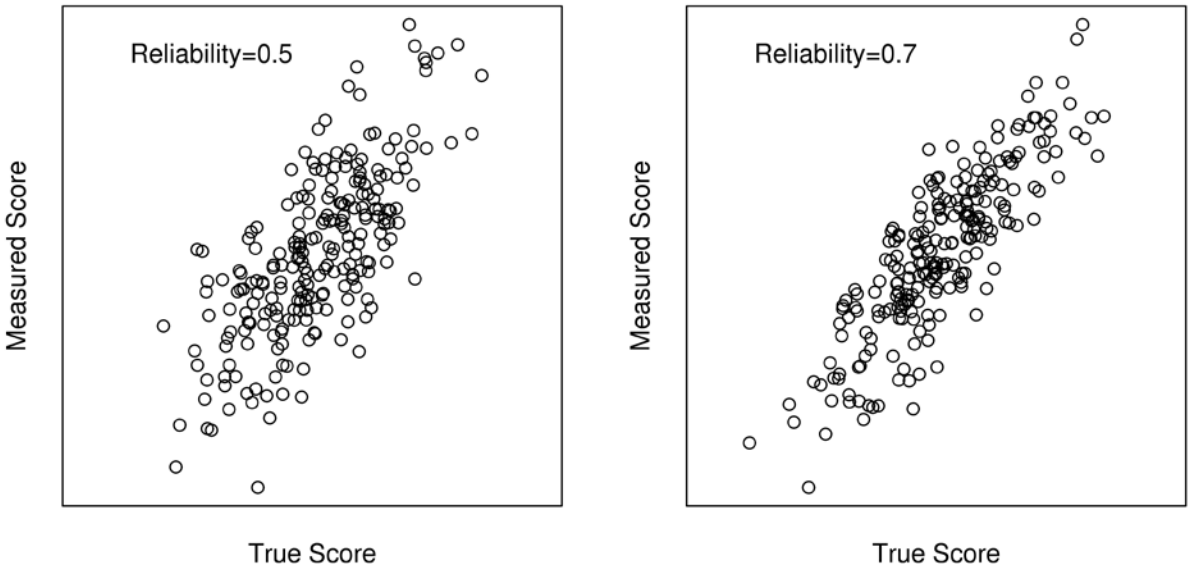
	No Minimum	**≥30 cases**	**≥50 cases**	**≥100 cases**	**≥150 cases**
No. of participants	233	186	156	101	53
Reliability	44.6%	51.7%	56.1%	60.9%	68.0%
95% CrI	(34.6%-54.1%)	(41.3%-61.4%)	(45.2%-65.6%)	(49.0%-71.2%)	(53.6%-79.7%)

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

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 Measured Score$_i$=True Score$_i$ + e_i where $i = 1,2,\ldots,233$ indicates the 233 participants and where True Score$_i$ and e_i both follow normal distributions. The standard deviations of the normal distributions were chosen such that the measure (score) has a reliability of 0.50 on the left figure and reliability of 0.70 on the right figure. Each figure has true score along the x-axis, and the estimated (measured) value of this true score along the y-axis. With a decreasing measurement error of the score (as is the case with increase in the number of cases per participant), the correlation between the true and measured values of the score increases, and thus also, equivalently, the reliability increases because reliability can be expressed as a square of this correlation (Pearson correlation). Although a high reliability of 0.70 shows a very close correlation between true and measured scores, a more moderate reliability of 0.50 still visualizes a strong association (correlation) between the true and measured values of the score.



2b1. VALIDITY TESTING

Note*: Current guidance for composite measure evaluation states that validity should be demonstrated for the composite performance measure score. If not feasible for initial endorsement, acceptable alternatives include assessment of content or face validity of the composite OR demonstration of validity for each component. Empirical validity testing of the composite measure score is expected by the time of endorsement maintenance.*

2b1.1. What level of validity testing was conducted?

☐ **Critical data elements** (*data element validity must address ALL critical data elements*)

☐ **Composite performance measure score**

☐ **Empirical validity testing**

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

☒ **Validity testing for component measures** *(check all that apply)*

Note*: applies to ALL component measures, unless already endorsed or are being submitted for individual endorsement.*

☐ **Endorsed (or submitted) as individual performance measures**

☒ **Critical data elements** (*data element validity must address ALL critical data elements*)

☐ **Empirical validity testing of the component measure score(s)**

☐ **Systematic assessment of face validity of <u>component measure score(s)</u> as an indicator** of quality or resource use (*i.e., is an accurate reflection of performance on quality or resource use and can distinguish good from poor performance*)

2b1.2. For each level of testing checked above, describe the method of validity testing and what it tests (*describe the steps—do not just name a method; what was tested, e.g., accuracy of data elements compared to authoritative source, relationship to another measure as expected; what statistical analysis was used)*
When data arrive at the data warehouse, they are checked carefully for logical inconsistencies, missing required fields, and parent/child variable relationship violations. Any inconsistencies or violations are communicated to participants in the detailed Data Quality Report that is generated automatically following each harvest file submission. Upon receipt of the Data Quality Report, participants are given an opportunity to correct the data, which substantially improves the quality and completeness of the data submitted for analysis. If the data inconsistencies are not changed by the participant prior to harvest close, the data warehouse performs consistency edits and/or parent/child edits on the data in order for them to be analyzable. Participants are informed of such edits to their data in the Data Quality Report.

Since 2010, the STS has contracted with Telligen (formerly IFMC) and, most recently, Cardiac Registry Support, LLC (CRS) to conduct audits of the STS General Thoracic Surgery Database on the Society´s behalf to evaluate the accuracy, consistency and comprehensiveness of data collection, which has validated the integrity of the data. Currently, auditors validate case inclusion and 15 lobectomy and 5 esophagectomy cancer cases are randomly chosen for review of 39 individual data elements. The auditors abstract each designated medical record to validate data elements previously submitted to the STS data warehouse. Agreement rates are calculated for each of the 39 elements as well as for an overall agreement rate. Five sites were randomly selected for the first audit, which took place in 2010. In 2016, 25 sites were audited.

2b1.3. What were the statistical results from validity testing? (*e.g., correlation; t-test*)
STS audited 10% of participants in the General Thoracic Surgery Database in 2016 using an independent auditing firm (CRS). The sites were randomly selected and audited for data completeness and accuracy. Auditors compared case logs at each facility and cases submitted to the STS GTSD to assess completeness of data submission. There was consistent agreement across all participants for data completeness. Data accuracy was assessed by reabstraction of 15 randomly chosen lobectomy cancer cases and 5 esophagectomy cancer cases, comparing 39 data elements in the medical chart with the data file submitted to the STS GTSD. The agreement rate was 96.78% for overall data accuracy in 2016, with a range in agreement from 94.3% to 99.0%.

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

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

CATEGORY	FIELD_NAME	NUM	DEN	Agreement
PRE-OPERATIVE EVALUATION	OVERALL_ALL_FIELDS	6455	6738	95.80%
PRE-OPERATIVE EVALUATION	Admission Date	497	500	99.40%
PRE-OPERATIVE EVALUATION	Prior Cardiothoracic Surgery	488	500	97.60%
PRE-OPERATIVE EVALUATION	Pre-Op Chemo-Current Malignancy	489	500	97.80%
PRE-OPERATIVE EVALUATION	Pre-Op Thoracic Radiation Therapy	489	500	97.80%

PRE-OPERATIVE EVALUATION	Diabetes	413	423	97.64%
PRE-OPERATIVE EVALUATION	Diabetes Therapy	68	82	82.93%
PRE-OPERATIVE EVALUATION	Cigarette Smoking	489	500	97.80%
PRE-OPERATIVE EVALUATION	Pulmonary Function Tests Performed	419	423	99.05%
PRE-OPERATIVE EVALUATION	FEV1 Predicted	316	414	76.33%
PRE-OPERATIVE EVALUATION	Zubrod Score	491	500	98.20%
PRE-OPERATIVE EVALUATION	Lung Cancer	420	423	99.29%
PRE-OPERATIVE EVALUATION	Clinical Staging Method-Lung-EBUS	408	419	97.37%
PRE-OPERATIVE EVALUATION	Clinical Staging Method-Lung-PET or PET/CT	397	419	94.75%
PRE-OPERATIVE EVALUATION	Lung Cancer Tumor Size-T	377	419	89.98%
PRE-OPERATIVE EVALUATION	Lung Cancer Nodes-N	409	419	97.61%
PRE-OPERATIVE EVALUATION	Esophageal Cancer	77	77	100.00%
PRE-OPERATIVE EVALUATION	Clinical Staging Method-Esophageal-EUS	69	75	92.00%
PRE-OPERATIVE EVALUATION	Esophageal Cancer Tumor-T	68	72	94.44%
PRE-OPERATIVE EVALUATION	Clinical Diagnosis of Nodal Involvement	71	73	97.26%
DIAGNOSIS AND PROCEDURES	OVERALL_ALL FIELDS	4842	4978	97.27%
DIAGNOSIS AND PROCEDURES	Category of Disease-Primary	479	499	95.99%
DIAGNOSIS AND PROCEDURES	Date of Surgery	498	500	99.60%
DIAGNOSIS AND PROCEDURES	Procedure Start Time	493	500	98.60%
DIAGNOSIS AND PROCEDURES	Procedure End Time	482	500	96.40%
DIAGNOSIS AND PROCEDURES	ASA Classification	487	500	97.40%
DIAGNOSIS AND PROCEDURES	Procedure	500	500	100.00%
DIAGNOSIS AND PROCEDURES	Patient Disposition	491	500	98.20%
DIAGNOSIS AND PROCEDURES	Pathologic Staging-Lung Cancer-T	405	419	96.66%
DIAGNOSIS AND PROCEDURES	Pathologic Staging-Lung Cancer-N	411	419	98.09%
DIAGNOSIS AND PROCEDURES	Lung Cancer-Number of Nodes	385	419	91.89%
DIAGNOSIS AND PROCEDURES	Pathologic Staging-Esophageal Cancer-T	69	74	93.24%
DIAGNOSIS AND PROCEDURES	Pathologic Staging-Esophageal Cancer-N	73	74	98.65%
DIAGNOSIS AND PROCEDURES	Esophageal Cancer-Number of Nodes	69	74	93.24%
POST-OPERATIVE EVENTS	OVERALL_ALL FIELDS	1487	1500	99.13%
POST-OPERATIVE EVENTS	Unexpected Return to OR	493	500	98.60%
POST-OPERATIVE EVENTS	Pneumonia	494	500	98.80%
POST-OPERATIVE EVENTS	Initial Vent Support >48 Hours	500	500	100.00%
DISCHARGE	OVERALL_ALL FIELDS	1935	1993	97.09%
DISCHARGE	Discharge Date	499	500	99.80%
DISCHARGE	Discharge Status	490	500	98.00%

DISCHARGE	Readmission within 30 Days of Discharge	484	493	98.17%
DISCHARGE	Status 30 Days After Surgery	462	500	92.40%
	OVERALL_ALL FIELDS	**14719**	**15209**	**96.78%**

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

2b2. EXCLUSIONS ANALYSIS
Note: *Applies to the composite performance measure, as well all component measures unless they are already endorsed or are being submitted for individual endorsement.*
NA ☐ no exclusions — *skip to section 2b4*

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

We excluded patients with missing data for age, sex, or discharge mortality status. In addition we excluded patients with non-elective status, occult or stage 0 tumors, 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.

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

There were 183 (0.7%) occult or stage 0 tumors, 8 (0.03%) ASA VI, and 337 (1.3%) non-elective status patients, resulting in the overall exclusion of 2.1% (528 of 24,912 patient records). Impact of these exclusions on the performance measure is negligible due to the small proportion of cases excluded.

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

For the measure to consistently quantify the surgical quality of 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, occult or stage 0 tumors, or American Society of Anesthesiologists class VI.

2b3. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES
Note: *Applies to all outcome or resource use component measures, unless already endorsed or are being*

submitted for individual endorsement.
If not an intermediate or health outcome, or PRO-PM, or resource use measure, skip to section 2b4.

2b3.1. What method of controlling for differences in case mix is used? *(check all that apply)*
☐ **Endorsed (or submitted) as individual performance measures**
☐ **No risk adjustment or stratification**
☒ **Statistical risk model with risk factors**
☐ **Stratification by risk categories**
☐ **Other,**

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

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.

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

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

Covariates in this model were selected a priori based on a combination of literature review and expert group consensus, and as described in 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.

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

☐ **Internal data analysis**
☒ **Other (please describe)**

Expert group consensus

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

Estimated odds ratios are summarized in the table below.

	Operative Mortality		Major Morbidity	
Variable	**OR (95% CI)**	**p-value**	**OR (95% CI)**	**p-value**
Age, yrs, (per 1 yr increase)	1.043 (1.027, 1.058)	<.0001	1.011 (1.006, 1.017)	<.0001
Male	1.372 (1.081, 1.743)	0.0094	1.377 (1.252, 1.514)	<.0001
Body Mass Index (kg/m2), (per 1 unit increase)	0.958 (0.937, 0.98)	0.0002	0.986 (0.978, 0.994)	0.0007
Hypertension	1.471 (1.106, 1.955)	0.0079	0.986 (0.889, 1.095)	0.7936
Steroid therapy	1.419 (0.844, 2.387)	0.1866	1.027 (0.797, 1.322)	0.839
Congestive heart failure	1.611 (1.004, 2.585)	0.0483	1.202 (0.942, 1.535)	0.1395
Coronary artery disease	1.308 (1.007, 1.698)	0.0443	1.286 (1.150, 1.438)	<.0001
Peripheral vascular disease	1.738 (1.298, 2.328)	0.0002	1.248 (1.085, 1.435)	0.0019
Reoperation	1.328 (0.894, 1.975)	0.1604	1.110 (0.926, 1.331)	0.2583
Preoperative chemotherapy within 6 months	1.229 (0.791, 1.911)	0.3592	1.268 (1.065, 1.509)	0.0075
Cerebrovascular disease	1.062 (0.744, 1.514)	0.7409	1.116 (0.955, 1.304)	0.1674
Diabetes mellitus	1.026 (0.775, 1.358)	0.8591	0.968 (0.858, 1.091)	0.5888
Renal failure	1.695 (0.873, 3.29)	0.119	1.387 (0.986, 1.95)	0.0604
Dialysis	4.110 (1.761, 9.596)	0.0011	1.005 (0.535, 1.888)	0.9885
Past smoker	1.172 (0.774, 1.776)	0.4533	1.522 (1.272, 1.821)	<.0001
Current smoker	1.411 (0.889, 2.238)	0.1441	2.168 (1.790, 2.627)	<.0001

FEV in 1 second percent of predicted (per 1 unit increase)	0.991 (0.985, 0.997)	0.0028	0.987 (0.985, 0.99)	<.0001
Zubrod score (per 1 unit increase)	1.233 (0.895, 1.699)	0.1997	1.182 (1.030, 1.355)	0.0172
Squared Zubrod score (per 1 unit increase)	1.039 (0.922, 1.17)	0.5295	1.021 (0.962, 1.083)	0.5003
ASA Class (per 1 unit increase)	2.160 (0.383, 12.181)	0.3828	1.127 (0.595, 2.137)	0.7139
Squared ASA Class (per 1 unit increase)	0.909 (0.691, 1.196)	0.4952	1.032 (0.931, 1.144)	0.5532
Pathologic stage I	1.216 (0.910, 1.626)	0.1867	1.200 (1.068, 1.349)	0.0022
Pathologic stage II	1.660 (1.199, 2.298)	0.0022	1.142 (0.984, 1.325)	0.0797
Pathologic stage IV	1.575 (0.686, 3.615)	0.2841	1.222 (0.862, 1.733)	0.2593
Year of operation (per 1 yr increase)	0.916 (0.797, 1.053)	0.2188	0.925 (0.874, 0.978)	0.0065

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

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

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

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

Continuous variables were evaluated with respect to linearity of effect and 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.

Provide the statistical results from testing the approach to controlling for differences in patient characteristics (case mix) below.
If stratified, skip to 2b3.9

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

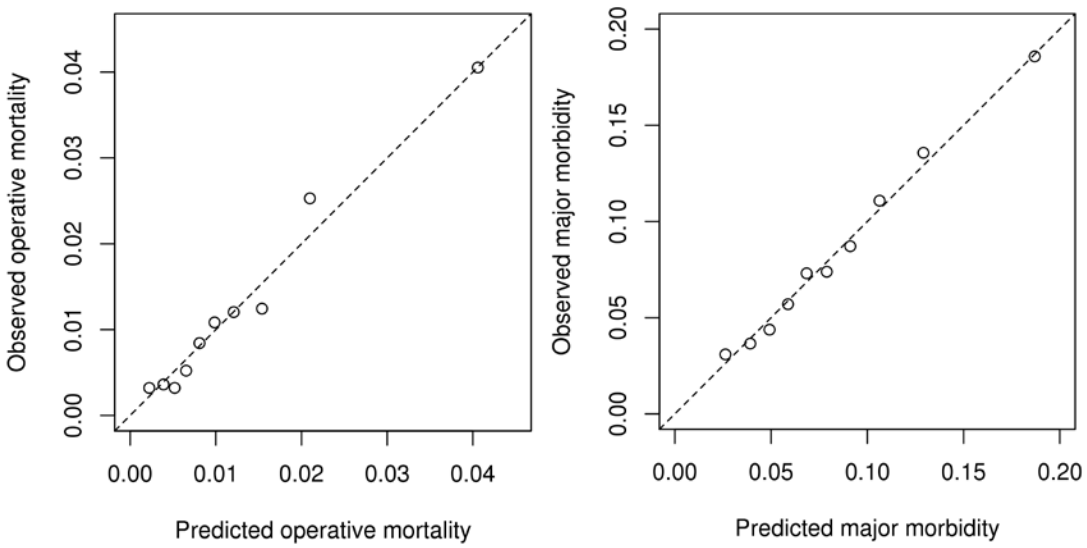
Operative mortality model: C-statistic is 0.731. Major morbidity model: C-statistic is 0.667.

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

Operative mortality model: Hosmer and Lemeshow Goodness-of-Fit Test p-value=0.47 (Chi-Square=7.65, df=8). Major morbidity model: Hosmer and Lemeshow Goodness-of-Fit Test p-value=0.44 (Chi-Square=7.95, df=8).

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

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



2b3.9. Results of Risk Stratification Analysis:
n/a

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

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

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

2b4. IDENTIFICATION OF STATISTICALLY SIGNIFICANT & MEANINGFUL DIFFERENCES IN PERFORMANCE
Note: *Applies to the composite performance measure.*

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

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).

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

Among participants with at least 30 cases over 3 years, 93.1% of participants have received 2 stars, and the remaining participants have received either 1 or 3 stars.

January 1, 2014 through December 31, 2016

	All Participants	**Participants N≥ 30**
Category	**Number of Participants, %**	**Number of Participants, %**
1-star	6, 2.6%	6, 3.2%
2-star	217, 93.1%	170, 91.4%
3-star	10, 4.3%	10, 5.4%

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

The 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.

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

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

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

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

factors in the risk adjustment model. However, if comparability is not demonstrated for measures with more than one set of specifications/instructions, the different specifications (e.g., for medical records vs. claims) should be submitted as separate measures.

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

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

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

2b6. MISSING DATA ANALYSIS AND MINIMIZING BIAS
Note: *Applies to the overall composite measure.*

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

The quality of data in the STS General Thoracic Surgery Database has been improving. We managed the missing data with imputation. Missing body mass index (BMI) values (1%) were imputed utilizing the median of the observed BMI values. Missing FEV1 (3.4%) was imputed to the median within the smoking status categories. Missing pathologic stage (3.1%) was imputed to its mode (stage I). For binary risk factors, missing values were considered as indicating absence of the risk factor.

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

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

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

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

2c. EMPIRICAL ANALYSIS TO SUPPORT COMPOSITE CONSTRUCTION APPROACH
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 if there is no empirical analysis.*

2d1. 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.

2d1.1 Describe the method used (*describe the steps—do not just name a method; what statistical analysis was used; if no empirical analysis, provide justification)*

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.

2d1.2. What were the statistical results obtained from the analysis of the components? (e.g., *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*)

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%)

Among 186 hospitals with at least 30 lobectomies.

2d1.3. What is 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? (i.*e., 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)*

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

2d2. 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

2d2.1 Describe the method used (*describe the steps—do not just name a method; what statistical analysis was used; if no empirical analysis, provide justification)*

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.

2d2.2. What were the statistical results obtained from the analysis of the aggregation and weighting rules? (e.g., *results of sensitivity analysis of effect of different aggregations and/or weighting rules; if no empirical analysis, identify the aggregation and weighting rules that were considered and the pros and cons of each*)

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.

2d2.3. What is your interpretation of the results in terms of demonstrating the aggregation and weighting rules are consistent with the described quality construct? (i.*e., 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*)

This weighting was consistent with our expert panel's clinical assessment of each domain's relative importance.

Statistical Model

For the i-th of n_j patients at the j-th participant ($j = 1, 2, \ldots, N$), let Y_{1ji} be a binary indicator of operative mortality status (0=alive, 1=dead), let Y_{2ji} be an indicator of major complications (0 = none, 1 = at least one), and let $\boldsymbol{x}_{ji} = (x_{1ji}, x_{2ji}, \ldots, x_{qji})$ be a set of numerically encoded patient baseline characteristics (e.g. age in years; binary risk factors coded as 0=absent, 1=present, etc.). Let $\pi_{kji} = \Pr(Y_{kji} = 1|x_{ji})$ denote the probability of the occurrence of the k-th endpoint where $k = 1$ refers to mortality and $k = 2$ refers to complications. The associations of x_{ji} with Y_{1ji} and Y_{2ji} are assumed to be described by a bivariate random effects logistic regression model with normally distributed hospital-specific random intercept parameters. In particular, we assume:

$$\begin{aligned}
&\text{(operative mortality)} \quad \log\left(\frac{\pi_{1ji}}{1-\pi_{1ji}}\right) = \alpha_{1j} + \boldsymbol{x}'_{ji}\boldsymbol{\beta}_1 \\
&\text{(major complication)} \quad \log\left(\frac{\pi_{2ji}}{1-\pi_{2ji}}\right) = \alpha_{2j} + \boldsymbol{x}'_{ji}\boldsymbol{\beta}_2 \\
&\text{(random effects)} \quad (\alpha_{1j}, \alpha_{2j}) \overset{\text{iid}}{\sim} N(\boldsymbol{\mu}, \boldsymbol{\Sigma}),
\end{aligned}$$

where $\boldsymbol{\beta}_1 = (\beta_{11}, \boldsymbol{\beta}_{12}, \ldots, \beta_{1q})$ denotes a set of unknown regression coefficients relating covariates to mortality, $\beta_2 = (\beta_{21}, \beta_{22}, \ldots, \beta_{2q})$ denotes a set of unknown regression coefficients relating covariates to major complication, $(\alpha_{1j}, \alpha_{2j})$ denote a set of normally distributed hospital-specific random effect parameters, and $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes a bivariate normal distribution with mean $\boldsymbol{\mu} = (\mu_1, \mu_2)$ and covariance $\boldsymbol{\Sigma} = (\sigma_{11}, \sigma_{12}, \sigma_{21}, \sigma_{22})$. Conditional on π_{1ji} and π_{2ji}, the variables Y_{1ji} and Y_{2ji} are assumed to be distributed as two independent Bernoulli variables with parameters π_{1ji} and π_{2ji}, respectively. That is:

$$\Pr(Y_{1ji} = y_{1ji}, Y_{2ji} = y_{2ji}|\pi_{1ji}, \pi_{2ji}) = \prod_{k=1}^{2} \pi_{kji}^{y_{kji}}(1 - \pi_{kji})^{1-y_{kji}}.$$

Outcomes of patients at different participants are assumed to be statistically independent, and outcomes of patients at the same participant are assumed to be conditionally independent given $(\alpha_{1j}, \alpha_{2j})$. The assumption that Y_{1ji} and Y_{2ji} are conditionally independent given π_{1ji} and π_{2ji} is likely to be violated in practice but is made in order to facilitate computation. Although the model assumes *conditional* independence between Y_{1ji} and Y_{2ji}, the model does not assume *marginal* independence between these two variables, as the underlying probabilities π_{1ji} and π_{2ji} depend on random effects parameters which account for within-hospital correlation.

Definition of Risk-Adjusted Rates

Based on this model, the j-th participant's risk-adjusted rates of operative mortality and major complications were defined as

$$\text{(operative mortality)} \qquad \theta_{1j} = \frac{\sum_{i=1}^{n_j} \text{expit}(\alpha_{1j} + \boldsymbol{x}'_{ji}\boldsymbol{\beta}_1)}{\sum_{i=1}^{n_j} \text{expit}(\mu_1 + \boldsymbol{x}'_{ji}\boldsymbol{\beta}_1)} \times \overline{Y}_1$$

$$\text{(major complication)} \qquad \theta_{2j} = \frac{\sum_{i=1}^{n_j} \text{expit}(\alpha_{2j} + \boldsymbol{x}'_{ji}\boldsymbol{\beta}_2)}{\sum_{i=1}^{n_j} \text{expit}(\mu_2 + \boldsymbol{x}'_{ji}\boldsymbol{\beta}_2)} \times \overline{Y}_2$$

where $\overline{Y}_1$ denotes the overall aggregate observed rate of operative mortality in the study sample and $\overline{Y}_2$ denotes the overall aggregate observed rate of major complication in the study sample.

Definition of Composite Score

The overall composite score of the j-th participant was defined as

$$\theta_j = w(1 - \theta_{1j}) + (1 - w)(1 - \theta_{2j})$$

where $w = (1/\sigma_1)/(1/\sigma_1 + 1/\sigma_2)$ and σ_k denotes the standard deviation of the θ_{kj}'s across participants, $k = 1, 2$.

Estimation

Model parameters were estimated in a Bayesian framework by specifying a prior probability distribution for the unknown model parameters $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$, $\boldsymbol{\mu}$, and $\boldsymbol{\Sigma}$. Because our prior knowledge was limited, we specified a vague proper prior distribution that consisted of independent normal distributions for the elements of $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$, and $\boldsymbol{\mu}$, and an inverse Wishart distribution for $\boldsymbol{\Sigma}$. Posterior means and credible intervals were calculated using Markov Chain Monte Carlo (MCMC) simulations as implemented in OpenBUGS version 3.2.2 software. Posterior summaries were calculated by generating 50,000 sets of simulated parameter values after a long burn-in period to ensure convergence and then thinning the sample to arrive at a final set of 5,000 iterations. The parameter θ_j was estimated as $\hat{\theta}_j = \sum_{l=1}^{5000} \theta_j^{(l)} \Big/ 5000$, where $\theta_j^{(l)}$ denotes the simulated values of θ_j at the l-th iteration of the MCMC procedure. A 95% Bayesian credible interval was obtained by calculating the 125th lowest and 125th highest values of θ_j across the 5000 simulated values.

bout the quality of care
ity to pass this criterion and

sistently within and across
sures Format (HQMF) and

ains current detailed
linking to a home page or to

e eMeasure authoring tool
plain-language description

ble) must be attached. (Excel

ls, questionnaires, scales,

ls, questionnaires,

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

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

No

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

n/a

S.4. **Numerator Statement** *(Brief, narrative description of the measure focus or what is being measured about the target population, i.e., cases from the target population with the target process, condition, event, or outcome) DO NOT include the rationale for the measure.*
IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm (S.14).
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.

S.5. **Numerator Details** *(All information required to identify and calculate the cases from the target population with the target process, condition, event, or outcome such as definitions, time period for data collection, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
IF an OUTCOME MEASURE, describe how the observed outcome is identified/counted. Calculation of the risk-adjusted outcome should be described in the calculation algorithm (S.14).
Number of patients undergoing elective 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

S.6. **Denominator Statement** *(Brief, narrative description of the target population being measured)*
Number of patients greater than or equal to 18 years of age undergoing elective lobectomy for lung cancer

S.7. **Denominator Details** *(All information required to identify and calculate the target population/denominator such as definitions, time period for data collection, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b.)*
IF an OUTCOME MEASURE, describe how the target population is identified. Calculation of the risk-adjusted outcome should be described in the calculation algorithm (S.14).
1. Lung cancer (LungCancer - STS GTS Database, v 2.2, sequence number 830) is marked "yes" and Category of Disease – Primary (CategoryPrim - STS GTS Database, v 2.2, sequence number 1300) is marked as one of the following:

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

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

Thoracoscopy, surgical; with lobectomy (32663)
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

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

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

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

S.11. **Risk Adjustment Type** (Select type. Provide specifications for risk stratification in measure testing attachment)
Statistical risk model
If other:

S.12. **Type of score:**
Rate/proportion
If other:

S.13. **Interpretation of Score** *(Classifies interpretation of score according to whether better quality is associated with a higher score, a lower score, a score falling within a defined interval, or a passing score)*
Better quality = Lower score

S.14. **Calculation Algorithm/Measure Logic** (*Diagram or describe the calculation of the measure score as an ordered sequence of steps including identifying the target population; exclusions; cases meeting the target process, condition, event, or outcome; time period for data, aggregating data; risk adjustment; etc.)*
Target population is patients 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.

S.15. **Sampling** *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*
IF an instrument-based performance measure (e.g., PRO-PM), identify whether (and how) proxy responses are allowed.
n/a

S.16. **Survey/Patient-reported data** *(If measure is based on a survey or instrument, provide instructions for data collection and guidance on minimum response rate.)*
Specify calculation of response rates to be reported with performance measure results.
n/a

S.17. **Data Source** *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*
If other, please describe in S.18.
Other, Registry Data

S.18. **Data Source or Collection Instrument** *(Identify the specific data source/data collection instrument (e.g. name of database, clinical registry, collection instrument, etc., and describe how data are collected.)*
IF instrument-based, identify the specific instrument(s) and standard methods, modes, and languages of administration.
STS General Thoracic Surgery Database, Version 2.3

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

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

S.21. **Care Setting** *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*
Inpatient/Hospital
If other:

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

2. Validity – See attached Measure Testing Submission Form

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

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

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

3. Feasibility

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

3a. Byproduct of Care Processes
For clinical measures, the required data elements are routinely generated and used during care delivery (e.g., blood pressure, lab test, diagnosis, medication order).

3a.1. **Data Elements Generated as Byproduct of Care Processes.**
Generated 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-9 codes on

claims), Abstracted from a record by someone other than person obtaining original information (e.g., chart abstraction for quality measure or registry)
If other:

3b. Electronic Sources

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

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

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

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

3c. Data Collection Strategy

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

3c.1. **Required for maintenance of endorsement. Describe difficulties (as a result of testing and/or operational use of the measure) regarding data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, time and cost of data collection, other feasibility/implementation issues.**
IF instrument-based, consider implications for both individuals providing data (patients, service recipients, respondents) and those whose performance is being measured.
Missing data are sought by the DCRI from participants when the data are initially sent to DCRI for analysis.

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

No individual patient identifiers are collected by the DCRI.

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

Other fees:
STS General Thoracic Surgery Database participant surgeons pay an annual participant fee of $550 or $700, depending on whether the participant is an STS member or not. STS membership thus provides surgeons with a 21% discount on the non-member database participation fee.

3c.2. **Describe any fees, licensing, or other requirements to use any aspect of the measure as specified** *(e.g., value/code set, risk model, programming code, algorithm)*.
See 3c.1

4. Usability and Use

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

4a. Accountability and Transparency

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

4.1. Current and Planned Use

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

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

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

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

See 4a1.2

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

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

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

As of July 2017, General Thoracic Surgery Database (GTSD) participants were included in the Public Reporting initiative and more than 250 participants currently consent to report outcomes publicly on the STS website. This includes discharge mortality rate and median postoperative length of stay for lobectomy procedures for lung cancer, including scores and star ratings for the Lobectomy for Lung Cancer Composite Measure in addition to its domains of 1) absence of mortality, and 2) absence of major complication. Participant outcomes are published alongside GTSD overall outcomes and National Inpatient Sample (NIS) outcomes.

-ACSD public reporting online may be found here: http://publicreporting.sts.org/acsd
-CHSD public reporting online may be found here: http://publicreporting.sts.org/chsd

-GHSD public reporting online may be found here: http://publicreporting.sts.org/gtsd

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

4a2.1.1. **Describe how performance results, data, and assistance with interpretation have been provided to those being measured or other users during development or implementation.**
How many and which types of measured entities and/or others were included? If only a sample of measured entities were included, describe the full population and how the sample was selected.
STS's combined mortality and morbidity model for pulmonary resection for lung cancer is important and appropriate for public reporting for the following reasons:
1.) within the broad category of lung cancer resections, lobectomy is the single most common major procedure that a thoracic surgeon performs;
2.) these procedures are therefore useful and appropriate to use as a benchmark for performance by general thoracic surgery programs. By providing surgeons and teams with risk-adjusted results, they can identify how they are performing compared with other programs in the STS General Thoracic Database, which generally includes the top thoracic programs in the nation. This will assist them in focusing performance improvement efforts. Also, when publicly reported, the outcomes for these common procedures provide patients and their families with comparative performance information to aid in selection of a provider;
3.) major morbidity is relatively common after lung resection; however, although mortality is rare, it should be captured as well in an outcome measure, thereby identifying ALL adverse events after lung resection;
4.) this measure is reported in an easy to understand format which summarizes the results of all participants who were included in the analysis. The participant's score is illustrated graphically in relation to the 25th, 50th and 75th percentiles of the distribution across participants, and is accompanied by the 95% Bayesian credible interval. Surgeons easily grasp this result and the visual display powerfully shows them just where they perform compared to their peers on a bi-annual basis. In addition, these risk-adjusted results allow surgeons to 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.

4a2.1.2. **Describe the process(es) involved, including when/how often results were provided, what data were provided, what educational/explanatory efforts were made, etc.**
See 4a2.1.1

4a2.2.1. **Summarize the feedback on measure performance and implementation from the measured entities and others described in 4d.1.**
Describe how feedback was obtained.
The general thoracic surgeons from across the U.S. who comprise the STS General Thoracic Surgery Task Force meet periodically to discuss the participant reports and to consider potential enhancements to the GTSD. Additions/clarifications to the data collection form and to the content/format of the participant reports are discussed and implemented as appropriate.

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

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

4a2.2.2. **Summarize the feedback obtained from those being measured.**
See 4a2.2.1

4a2.2.3. **Summarize the feedback obtained from other users**
Given the very recent launch of general thoracic public reporting, the STS has not yet received sufficient feedback from non-participants to be able to assess the impact of the public reporting initiative.

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

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

4b1. **Refer to data provided in 1b but do not repeat here. Discuss any progress on improvement (trends in performance results, number and percentage of people receiving high-quality healthcare; Geographic area and number and percentage of accountable entities and patients included.)**
If no improvement was demonstrated, what are the reasons? If not in use for performance improvement at the time of initial endorsement, provide a credible rationale that describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.
Operative mortality in the STS General Thoracic Surgery Database has decreased from 2.2% in the years 2002 to 2008 to 1.4% from 2012 to 2014. These data represent the highest quality lung cancer surgery in the United States. It is important to recognize that a large proportion of the general thoracic surgery in the US is not performed by general thoracic surgeons certified by the American Board of Thoracic Surgery. Results by STS General Thoracic Database participants, who are almost all ABTS certified, are generally superior to those of surgeons performing these procedures who do not participate in the GTSD, and who are often not ABTS certified.

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

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

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

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

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

4b2.1. **Please explain any unexpected findings (positive or negative) during implementation of this measure including unintended impacts on patients.**
We are not aware of any unexpected findings associated with implementation of this measure.

4b2.2. **Please explain any unexpected benefits from implementation of this measure.**
n/a

5. Comparison to Related or Competing Measures

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

5. Relation to Other NQF-endorsed Measures
Are there related measures (conceptually, either same measure focus or target population) or competing measures (conceptually both the same measure focus and same target population)? If yes, list the NQF # and title of all related and/or competing measures.
Yes

5.1a. **List of related or competing measures (selected from NQF-endorsed measures)**
1790 : Risk-Adjusted Morbidity and Mortality for Lung Resection for Lung Cancer

5.1b. **If related or competing measures are not NQF endorsed please indicate measure title and steward.**
n/a (measure #1790 is NQF endorsed, eligible for endorsement maintenance in this Surgery Project cycle)

5a. Harmonization of Related Measures
The measure specifications are harmonized with related measures;
OR
The differences in specifications are justified

5a.1. **If this measure conceptually addresses EITHER the same measure focus OR the same target population as NQF-endorsed measure(s):**
Are the measure specifications harmonized to the extent possible?
Yes

5a.2. **If the measure specifications are not completely harmonized, identify the differences, rationale, and impact on interpretability and data collection burden.**
Measure #1790 includes a broader range of lung resection procedures than the Lobectomy Composite, and therefore includes a larger number of cases and potentially provides performance data to more general thoracic surgeons. Of the two measures, only the Lobectomy Composite is currently publicly reported.

5b. Competing Measures
The measure is superior to competing measures (e.g., is a more valid or efficient way to measure);
OR
Multiple measures are justified.

5b.1. **If this measure conceptually addresses both the same measure focus and the same target population as NQF-endorsed measure(s):**
Describe why this measure is superior to competing measures (e.g., a more valid or efficient way to measure quality); OR provide a rationale for the additive value of endorsing an additional measure. (Provide analyses when possible.)
n/a

Appendix

A.1 Supplemental materials may be provided in an appendix. All supplemental materials (such as data collection instrument or methodology reports) should be organized in one file with a table of contents or bookmarks. If material pertains to a specific submission form number, that should be indicated. Requested information should be provided in the submission form and required attachments. There is no guarantee that supplemental materials will be reviewed.
Attachment **Attachment:** KozowerOBrienKosinski-et-al-2016-PIIS0003497515017531.pdf

Contact Information

Co.1 **Measure Steward (Intellectual Property Owner):** The Society of Thoracic Surgeons
Co.2 **Point of Contact:** Mark, Antman, mantman@sts.org, 312-202-5856-
Co.3 **Measure Developer if different from Measure Steward:** The Society of Thoracic Surgeons
Co.4 **Point of Contact:** Mark, Antman, mantman@sts.org, 312-202-5856-

Additional Information

Ad.1 **Workgroup/Expert Panel involved in measure development**

Provide a list of sponsoring organizations and workgroup/panel members' names and organizations. Describe the members' role in measure development. Members of the STS Task Force on Quality Initiatives provide surgical expertise as needed. The STS Workforce on National Databases meets at the STS Annual Meeting and reviews the measures on a yearly basis. Changes or updates to the measure will be at the recommendation of the Workforce.
Measure Developer/Steward Updates and Ongoing Maintenance Ad.2 **Year the measure was first released:** 2016 Ad.3 **Month and Year of most recent revision:** 01, 2016 Ad.4 **What is your frequency for review/update of this measure?** annually Ad.5 **When is the next scheduled review/update for this measure?** 01, 2018
Ad.6 **Copyright statement:** Ad.7 **Disclaimers:**
Ad.8 **Additional Information/Comments:**

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 #: 1790
Measure Title: Risk-Adjusted Morbidity and Mortality for Lung Resection for Lung Cancer
Measure Steward: The Society of Thoracic Surgeons
Brief Description of Measure: Percentage of patients greater than or equal to 18 years of age undergoing elective lung resection (Open or VATS wedge resection, segmentectomy, lobectomy, bilobectomy, sleeve lobectomy, pneumonectomy) for lung cancer who developed any of the following postoperative complications: reintubation, need for tracheostomy, initial ventilator support > 48 hours, ARDS, pneumonia, pulmonary embolus, bronchopleural fistula, unexpected return to the operating room, myocardial infarction or operative mortality (death during the index hospitalization, regardless of timing, or within 30 days, regardless of location).
Developer Rationale: Providing outcomes data to participating thoracic surgery sites allows benchmarking of practice group results against the STS national results and allows demonstration of improvement when QI efforts are undertaken. These outcomes data aid clinicians and patients in making informed clinical decisions and also enable them to compare risk-adjusted outcomes for quality improvement purposes.

Numerator Statement: Number of patients greater than or equal to 18 years of age undergoing elective lung resection (Open or VATS wedge resection, segmentectomy, lobectomy, bilobectomy, sleeve lobectomy, pneumonectomy) for lung cancer who developed any of the following postoperative complications: reintubation, need for tracheostomy, initial ventilator support > 48 hours, ARDS, pneumonia, pulmonary embolus, bronchopleural fistula, unexpected return to the operating room, myocardial infarction or operative mortality (death during the index hospitalization, regardless of timing, or within 30 days, regardless of location).
Denominator Statement: Number of patients greater than or equal to 18 years of age undergoing elective lung resection (Open or VATS wedge resection, segmentectomy, lobectomy, bilobectomy, sleeve lobectomy, pneumonectomy) for lung cancer
Denominator Exclusions: Patients were excluded if they had an extrapleural pneumonectomy, completion pneumonectomy, carinal pneumonectomy, occult carcinoma or benign disease on final pathology, or an urgent, emergent, or palliative operation. Furthermore, patients with missing age, sex, discharge mortality status, and predicted forced expiratory volume in 1 second were also excluded.

Measure Type: Outcome
Data Source: Other, Registry Data
Level of Analysis: Facility

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

Maintenance of Endorsement -- Preliminary Analysis

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

Criteria 1: Importance to Measure and Report

1a. Evidence

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

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

Summary of prior review in 2012

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

Changes to evidence from last review

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

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

Updates:

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

Question for the Committee:

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

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

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

1b. Gap in Care/Opportunity for Improvement and 1b. Disparities
Maintenance measures – increased emphasis on gap and variation

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

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

during January 1, 2012 through December 31, 2014. The distribution of hospital specific estimates of the SIR for morbidity and mortality is shown below.

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

Disparities

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

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

Questions for the Committee:

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

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

Committee pre-evaluation comments

Criteria 1: Importance to Measure and Report (including 1a, 1b, 1c)

1a. Evidence to Support Measure Focus

**Ample evidence from the STS registry is provided to calculate this measure.
** Evidence well supported
** Outcome measure with good data to support it
** This maintenance measure uses the well established Society for Thoracic Surgery risk adjusted database to evaluate the mortality and major morbidities after lung resection for lung cancer. Studies are cited that address the data integrity and utility of the measure.

1b. Performance Gap

**Gap is relatively small (3.5% good performers and 6.9% bad performers) but this measure is the most important outcome of surgery and therefore important for public accountability.
**Performance gap present
**There remains a performance gap for this existing measure
****A performance gap was demonstrated by the almost 5 fold difference between high-performing and low performing hospitals. Only racial disparities were addressed in the submission.

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability: Specifications and Reliability

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

Reliability

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

2a2. Reliability testing demonstrates if the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise enough to distinguish differences in performance across providers. For maintenance measures – less emphasis if no new testing data provided.

Validity

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

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

Composite measures only:

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

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

Evaluators: Michael Stoto, Zhenqiu Lin, Susan White

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

Evaluation A

Evaluation B

Evaluation C

Questions for the Committee regarding reliability:

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

Questions for the Committee regarding validity:

- *Do you have any concerns regarding the validity of the measure (e.g., exclusions, risk-adjustment approach, etc.)?*
- *The Scientific Methods Panel is satisfied with the validity analyses for the measure. Does the Committee think there is a need to discuss and/or vote on validity?*

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

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

Committee pre-evaluation comments

Criteria 2: Scientific Acceptability of Measure Properties (including all 2a, 2b, and 2c)

2a1. Reliability Specifications

**None

**High Reliability

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

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

2a2. Reliability Testing

**No

**No

**No

2b1. Validity Testing
**No
**Valid
**No
**The validity of the STS database has been evaluated in detail.

2b2-3. Other threats to validity
**Appropriately risk adjusted, and not unduly burdensome for those already participating in the STS registry. It will be nearly impossible to use this measure otherwise, however.
**Risk adjusted w adequate addressing of GTSDB lack of social risk factors
** I have no issues
**Appropriate exclusions have been made, primarily for low volume but high risk procedures that otherwise may skew the results.

Criterion 3. Feasibility
Maintenance measures – no change in emphasis – implementation issues may be more prominent

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

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

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

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

Committee pre-evaluation comments
Criteria 3: Feasibility

3a. Feasibility
**STS registry participation is expensive and burdensome, but offers a substantial return on investment.
**My only concern is penetrance of STS GTSDB. Historically the GTSDB is comprised of the highest TS performers.
**Participation in STS is costly, however there is widespread participation among facilities performing these procedures
**Although cost and other resources are required to participate in the STS registry, all but a few centers in the United States are currently participating.

Criterion 4: Usability and Use
Maintenance measures – increased emphasis – much greater focus on measure use and usefulness, including both impact/improvement and unintended consequences

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

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

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

Current uses of the measure

Publicly reported? ☒ **Yes** ☐ **No**

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

OR

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

Accountability program details

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

4a.2. Feedback on the measure by those being measured or others. Three criteria demonstrate feedback: 1) those being measured have been given performance results or data, as well as assistance with interpreting the measure results and data; 2) those being measured and other users have been given an opportunity to provide feedback on the measure performance or implementation; 3) this feedback has been considered when changes are incorporated into the measure

Feedback on the measure by those being measured or others:

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

Additional Feedback:

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

***Questions for the Committee**:*

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

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

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

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

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

Improvement results

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

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

Unexpected findings (positive or negative) during implementation

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

Potential harms

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

***Questions for the Committee**:*

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

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

Committee pre-evaluation comments

Criteria 4: Usability and Use

4a1. Use
**Closely followed by facilities and surgeons.
**Transparent
**No issues
**Participating institutions receive risk-adjusted reports has to their performance. The STS also has a public reporting task force that develops report cards for the consumer.

4b1. Usability
**Public reporting will increase attention to performance
**Usable
**No issues other than the cost to obtain the clinical data
**The institution level reports are designed to guide process improvement initiatives.

Criterion 5: Related and Competing Measures

Related or competing measures

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

Harmonization

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

Committee pre-evaluation comments

Criterion 5: Related and Competing Measures

Public and member comments

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

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

Evaluation A

Scientific Acceptability

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

Instructions:

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

Measure Number: 1790
Measure Title: Risk-Adjusted Morbidity and Mortality for Lung Resection for Lung Cancer

RELIABILITY

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

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

   ☒Yes (go to Question #2)

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

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

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

   ☒Yes (go to Question #4)

   ☐No, there is reliability testing information, but *not* using statistical tests and/or not for the

measure as specified OR there is no reliability testing (please explain below then go to Question #3)

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

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

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

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

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

   ☒**Yes (go to Question #5)**

   ☐No (go to Question #8)

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

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

   ☒**Yes (go to Question #6)**

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

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

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

   ☒**High (go to Question #8)**

   ☐Moderate (go to Question #8)

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

7. Was other reliability testing reported?

   ☐Yes (go to Question #8)

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

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

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

   ☐Yes (go to Question #9)

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

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

TIPS: For example: inter-abstractor agreement (ICC, Kappa); other accepted method with description of how it assesses reliability of the data elements
Answer no if: only assessed percent agreement; did not assess separately for all data elements (at least numerator, denominator, exclusions)

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

10. **RATING (data element)** – Based on the reliability statistic and scope of testing (number and representativeness of patients and entities), what is the level of certainty or confidence that the data used in the measure are reliable?
    TIPS: Consider the following: Is the test sample adequate to generalize for widespread implementation? Can data elements be collected consistently?
    ☐Moderate (if score-level testing was NOT conducted, rate Question #11: OVERALL RELIABILITY as MODERATE)
    ☐Low (if score-level testing was NOT conducted, rate Question #11: OVERALL RELIABILITY as LOW)
    ☐Insufficient (go to Question #11)

11. OVERALL RELIABILITY RATING

OVERALL RATING OF RELIABILITY taking into account precision of specifications and <u>all</u> testing results:

☒**High** **(NOTE: Can be HIGH <u>only if</u> score-level testing has been conducted)**
☐Moderate (NOTE: Moderate is the highest eligible rating if score-level testing has <u>not</u> been conducted)
☐Low (please explain below) [NOTE: Should rate LOW if you believe specifications are NOT precise, unambiguous, and complete]
☐Insufficient (please explain below) [NOTE: For most measure types, testing at both the score level and the data element level is <u>not</u> required]

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

VALIDITY

Assessment of Threats to Validity

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

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

   ☒Yes (go to Question #2)

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

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

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

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

   ☒No (go to Question #3)

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

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

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

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

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

   c. Any concerns regarding the risk-adjustment approach?

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

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

   ☐No (go to Question #4)

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

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

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

   ☒No (go to Question #5)

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

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

   ☒No (go to Question #6)

☐Not applicable (go to Question #6)

6. Analysis of potential threats to validity: Any concerns regarding missing data?
   ☐Yes (please explain below then go to Question #7)
   ☒**No (go to Question #7)**

Assessment of Measure Testing

7. Was empirical validity testing conducted using the measure as specified and appropriate statistical test?
   Answer no if: face validity; only refer to clinical evidence; only descriptive statistics; only describe process for data management/cleaning/computer programming; testing does not match measure specifications (i.e. data, eMeasure, level, setting, patients).
   ☒**Yes (go to Question #10)** [NOTE: If appropriate empirical testing has been conducted, then evaluation of face validity is not necessary. Go to Question #8 **only if** there is insufficient information provided to evaluate data element and score-level testing.]
   ☐No (please explain below then go to Question #8)

8. Was face validity systematically assessed by recognized experts to determine agreement on whether the computed performance measure score from the measure as specified can be used to distinguish good and poor quality?
   TIPS: Answer no if: focused on data element accuracy/availability/feasibility/other topics; the degree of consensus and any areas of disagreement not provided/discussed.
   ☐Yes (go to Question #9)
   ☐No (please explain below and rate Question #17: OVERALL VALIDITY as INSUFFICIENT)

9. **RATING (face validity)** - Do the face validity testing results indicate substantial agreement that the performance measure score from the measure as specified can be used to distinguish quality AND potential threats to validity are not a problem, OR are adequately addressed so results are not biased?
   ☐Yes (if a NEW measure, rate Question #17: OVERALL VALIDITY as MODERATE)
   ☐Yes (if a MAINTENANCE measure, do you agree with the justification for not conducting empirical testing? If no, rate Question #17: OVERALL VALIDITY as INSUFFICIENT; otherwise, rate Question #17: OVERALL VALIDITY as MODERATE)
   ☐No (please explain below and rate Question #17: OVERALL VALIDITY AS LOW)

10. Was validity testing conducted with computed performance measure scores for each measured entity?
    TIPS: Answer no if: one overall score for all patients in sample used for testing patient-level data.
    ☒**Yes (go to Question #11)**
    ☐No (please explain below and go to Question #13)

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

11. Was the method described and appropriate for assessing conceptually and theoretically sound hypothesized relationships?
    TIPS: For example: correlation of the performance measure score on this measure and other performance measures; differences in performance scores between groups known to differ on quality; other accepted method with description of how it assesses validity of the performance score

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

12. **RATING (measure score)** - Based on the measure score results (significance, strength) and scope of testing (number of measured entities and representativeness) and analysis of potential threats, what is the level of certainty or confidence that the performance measure scores are a valid indicator of quality?
☐High (go to Question #14)
☒**Moderate (go to Question #14)**
☐Low (please explain below then go to Question #13)
☐Insufficient

13. Was other validity testing reported?
☒**Yes (go to Question #14)**
☐No (please explain below and rate Question #17: OVERALL VALIDITY as LOW)

14. Was validity testing conducted with patient-level data elements?
TIPS: Prior validity studies of the same data elements may be submitted
☒**Yes (go to Question #15)**
☐No (please explain below and rate Question #17: OVERALL VALIDITY as INSUFFICIENT if no score-level testing was conducted, otherwise, rate Question #17: OVERALL VALIDITY based on score-level rating from Question #12)

15. Was the method described and appropriate for assessing the accuracy of ALL critical data elements? *NOTE that data element validation from the literature is acceptable.*
TIPS: For example: Data validity/accuracy as compared to authoritative source- sensitivity, specificity, PPV, NPV; other accepted method with description of how it assesses validity of the data elements.
Answer No if: only assessed percent agreement; did not assess separately for all data elements (at least numerator, denominator, exclusions)
☒**Yes (go to Question #16)**
☐No (please explain below and rate Question #16 as INSUFFICIENT)

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

16. **RATING (data element)** - Based on the data element testing results (significance, strength) and scope of testing (number and representativeness of patients and entities) and analysis of potential threats, what is the level of certainty or confidence that the data used in the measure are valid?
☒**Moderate (if score-level testing was NOT conducted, rate Question #17: OVERALL VALIDITY as MODERATE)**
☐Low (please explain below) (if score-level testing was NOT conducted, rate Question #17: OVERALL VALIDITY as LOW)

☐Insufficient (go to Question #17)

17. OVERALL VALIDITY RATING

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

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

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

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

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

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

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

FOR COMPOSITE MEASURES ONLY: Empirical analyses to support composite construction

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

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

☐High

☐Moderate

☐Low (please explain below)

☐Insufficient (please explain below)

Evaluation B

Scientific Acceptability

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

Instructions:

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

Measure Number: 1790
Measure Title: Risk-Adjusted Morbidity and Mortality for Lung Resection for Lung Cancer

RELIABILITY

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

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

    ☒Yes (go to Question #2)

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

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

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

    ☒Yes (go to Question #4)

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

13. Was **empirical VALIDITY testing** of patient-level data conducted?
    - ☐Yes (use your rating from data element validity testing – Question #16- under Validity Section)
    - ☐No (please explain below and rate Question #11: OVERALL RELIABILITY as INSUFFICIENT and proceed to the VALIDITY SECTION)

14. Was reliability testing conducted with computed performance measure scores for each measured entity?
    TIPS: Answer no if: only one overall score for all patients in sample used for testing patient-level data
    - ☒Yes (go to Question #5)
    - ☐No (go to Question #8)

15. Was the method described and appropriate for assessing the proportion of variability due to real differences among measured entities? *NOTE: If multiple methods used, at least one must be appropriate.*
    TIPS: Examples of appropriate methods include signal-to-noise analysis (e.g. Adams/RAND tutorial); random split-half correlation; other accepted method with description of how it assesses reliability of the performance score.
    - ☒Yes (go to Question #6)
    - ☐No (please explain below then go to Question #8)

16. **RATING (score level)** - What is the level of certainty or confidence that the performance measure scores are reliable?
    TIPS: Consider the following: Is the test sample adequate to generalize for widespread implementation? Do the results demonstrate sufficient reliability so that differences in performance can be identified?
    - ☐High (go to Question #8)
    - ☒Moderate (go to Question #8)
    - ☐Low (please explain below then go to Question #7)

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

17. Was other reliability testing reported?
    - ☒Yes (go to Question #8)
    - ☐No (rate Question #11: OVERALL RELIABILITY as LOW and proceed to the VALIDITY SECTION)

18. Was reliability testing conducted with patient-level data elements that are used to construct the performance measure?
    TIPS: Prior reliability studies of the same data elements may be submitted; if comparing abstraction to "authoritative source/gold standard" see Validity Section Question #15)
    - ☐Yes (go to Question #9)
    - ☒No (if there is score-level testing, rate Question #11: OVERALL RELIABILITY based on score-level rating from Question #6; otherwise, rate Question #11: OVERALL RELIABILITY as INSUFFICIENT. Then proceed to the VALIDITY SECTION)

    Developer mentions testing elements via random review (mentioned in Section 2b1.2 under validity). I think the audit is actually testing reliability and should be reported as a Kappa statistic. If the measure

developer is treating the auditing firm as the 'gold standard', then this could be considered a validity measure. As reported, it is hard to determine how they are treating the audit – other than including the description in the 'validity' section of the report.

19. Was the method described and appropriate for assessing the reliability of ALL critical data elements?
    TIPS: For example: inter-abstractor agreement (ICC, Kappa); other accepted method with description of how it assesses reliability of the data elements
    Answer no if: only assessed percent agreement; did not assess separately for all data elements (at least numerator, denominator, exclusions)
    ☐Yes (go to Question #10)
    ☐No (if no, please explain below and rate Question #10 as INSUFFICIENT)

20. **RATING (data element)** – Based on the reliability statistic and scope of testing (number and representativeness of patients and entities), what is the level of certainty or confidence that the data used in the measure are reliable?
    TIPS: Consider the following: Is the test sample adequate to generalize for widespread implementation? Can data elements be collected consistently?
    ☐Moderate (if score-level testing was NOT conducted, rate Question #11: OVERALL RELIABILITY as MODERATE)
    ☐Low (if score-level testing was NOT conducted, rate Question #11: OVERALL RELIABILITY as LOW)
    ☐Insufficient (go to Question #11)

11. OVERALL RELIABILITY RATING

OVERALL RATING OF RELIABILITY taking into account precision of specifications and <u>all</u> testing results:

☐High (NOTE: Can be HIGH <u>only if</u> score-level testing has been conducted)
☒Moderate (NOTE: Moderate is the highest eligible rating if score-level testing has <u>not</u> been conducted)
☐Low (please explain below) [NOTE: Should rate LOW if you believe specifications are NOT precise, unambiguous, and complete]
☐Insufficient (please explain below) [NOTE: For most measure types, testing at both the score level and the data element level is <u>not</u> required]

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

VALIDITY

Assessment of Threats to Validity

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

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

☒Yes (go to Question #2)

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

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

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

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

☒No (go to Question #3)

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

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

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

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

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

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

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

c. Any concerns regarding the risk-adjustment approach?

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

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

☐No (go to Question #4)

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

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

20. Analysis of potential threats to validity: Any concerns regarding ability to identify meaningful differences in performance or overall poor performance?
    ☒Yes (please explain below then go to Question #5)
    ☐No (go to Question #5)

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

21. Analysis of potential threats to validity: Any concerns regarding comparability of results if multiple data sources or methods are specified?
    ☐Yes (please explain below then go to Question #6)
    ☐No (go to Question #6)
    ☒Not applicable (go to Question #6)

22. Analysis of potential threats to validity: Any concerns regarding missing data?
    ☐Yes (please explain below then go to Question #7)
    ☒No (go to Question #7)

Assessment of Measure Testing

23. Was empirical validity testing conducted using the measure as specified and appropriate statistical test?
    Answer no if: face validity; only refer to clinical evidence; only descriptive statistics; only describe process for data management/cleaning/computer programming; testing does not match measure specifications (i.e. data, eMeasure, level, setting, patients).
    ☒Yes (go to Question #10) [NOTE: If appropriate empirical testing has been conducted, then evaluation of face validity is not necessary. Go to Question #8 **only if** there is insufficient information provided to evaluate data element and score-level testing.]
    ☐No (please explain below then go to Question #8)
    Agreement rates with auditor reported – see Question #8 under reliability.

24. Was face validity systematically assessed by recognized experts to determine agreement on whether the computed performance measure score from the measure as specified can be used to distinguish good and poor quality?
    TIPS: Answer no if: focused on data element accuracy/availability/feasibility/other topics; the degree of consensus and any areas of disagreement not provided/discussed.
    ☐Yes (go to Question #9)
    ☐No (please explain below and rate Question #17: OVERALL VALIDITY as INSUFFICIENT)

25. **RATING (face validity)** - Do the face validity testing results indicate substantial agreement that the performance measure score from the measure as specified can be used to distinguish quality AND potential threats to validity are not a problem, OR are adequately addressed so results are not biased?
    ☐Yes (if a NEW measure, rate Question #17: OVERALL VALIDITY as MODERATE)
    ☐ Yes (if a MAINTENANCE measure, do you agree with the justification for not conducting empirical testing? If no, rate Question #17: OVERALL VALIDITY as INSUFFICIENT; otherwise, rate Question #17: OVERALL VALIDITY as MODERATE)
    ☐No (please explain below and rate Question #17: OVERALL VALIDITY AS LOW)

26. Was validity testing conducted with computed performance measure scores for each measured entity?
    TIPS: Answer no if: one overall score for all patients in sample used for testing patient-level data.
    ☐Yes (go to Question #11)

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

I am interpreting 'measured entity' to be hospital for this measure. I do not see a hospital level assessment of validity on the computed score. I do see those results for the individual data elements for a sample of the measured entities.

27. Was the method described and appropriate for assessing conceptually and theoretically sound hypothesized relationships?

TIPS: For example: correlation of the performance measure score on this measure and other performance measures; differences in performance scores between groups known to differ on quality; other accepted method with description of how it assesses validity of the performance score

☐Yes (go to Question #12)

☐No (please explain below, rate Question #12 as INSUFFICIENT and then go to Question #14)

28. **RATING (measure score)** - Based on the measure score results (significance, strength) and scope of testing (number of measured entities and representativeness) and analysis of potential threats, what is the level of certainty or confidence that the performance measure scores are a valid indicator of quality?

☐High (go to Question #14)

☐Moderate (go to Question #14)

☐Low (please explain below then go to Question #13)

☐Insufficient

29. Was other validity testing reported?

☒Yes (go to Question #14)

☐No (please explain below and rate Question #17: OVERALL VALIDITY as LOW)

30. Was validity testing conducted with <u>patient-level data elements</u>?

TIPS: Prior validity studies of the same data elements may be submitted

☒Yes (go to Question #15)

☐No (please explain below and rate Question #17: OVERALL VALIDITY as INSUFFICIENT if <u>no</u> score-level testing was conducted, otherwise, rate Question #17: OVERALL VALIDITY based on score-level rating from Question #12)

Reported agreement rates with quality audits as validity measure. Agreement rates are high for most data elements (95% +)

31. Was the method described and appropriate for assessing the accuracy of ALL critical data elements? *NOTE that data element validation from the literature is acceptable.*

TIPS: For example: Data validity/accuracy as compared to authoritative source- sensitivity, specificity, PPV, NPV; other accepted method with description of how it assesses validity of the data elements.
Answer No if: only assessed percent agreement; did not assess separately for all data elements (at least numerator, denominator, exclusions)

☒Yes (go to Question #16)

☐No (please explain below and rate Question #16 as INSUFFICIENT)

32. **RATING (data element) -** Based on the data element testing results (significance, strength) and scope of testing (number and representativeness of patients and entities) and analysis of potential threats, what is the level of certainty or confidence that the data used in the measure are valid?

☒Moderate (if score-level testing was NOT conducted, rate Question #17: OVERALL VALIDITY as MODERATE)

☐Low (please explain below) (if score-level testing was NOT conducted, rate Question #17: OVERALL VALIDITY as LOW)

☐Insufficient (go to Question #17)

17. OVERALL VALIDITY RATING

OVERALL RATING OF VALIDITY taking into account the results and scope of all testing and analysis of potential threats.

☒High (NOTE: Can be HIGH only if score-level testing has been conducted)

☐Moderate (NOTE: Moderate is the highest eligible rating if score-level testing has NOT been conducted)

☐Low (please explain below) [NOTE: Should rate LOW if you believe that there are threats to validity and/or threats to validity were not assessed]

☐Insufficient (if insufficient, please explain below) [NOTE: For most measure types, testing at both the score level and the data element level is not required] [NOTE: If rating is INSUFFICIENT for all empirical testing, then go back to Question #8 and evaluate any face validity that was conducted, then reconsider this overall rating.]

FOR COMPOSITE MEASURES ONLY: Empirical analyses to support composite construction

What is the level of certainty or confidence that the empirical analysis demonstrates that the component measures add value to the composite and that the aggregation and weighting rules are consistent with the quality construct?

TIPS: Consider the following: Do the component measures fit the quality construct? Are the objectives of parsimony and simplicity achieved while supporting the quality construct?

☐High

☐Moderate

☐Low (please explain below)

☐Insufficient (please explain below)

Evaluation C

Scientific Acceptability

Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. **Measures must be judged to meet the subcriteria for both reliability and validity to pass this criterion.**

Instructions:

- Please complete this form for each measure you are evaluating.
- Please pay close attention to the skip logic directions.
- If you are unable to check a box, please highlight or shade the box for your response.
- You must answer the "overall rating" item for both Reliability and Validity. Also, be sure to answer the composite measure question at the end of the form if your measure is a composite.
- We have provided TIPS to help you answer the questions.
- We've designed this form to try to minimize the amount of writing that you have to do. That said, ***it is critical that you explain your thinking/rationale if you check boxes where we ask for an explanation*** (because this is a Word document, you can just add your explanation below the checkbox)**.** Feel free to add additional explanation, even if an explanation is not requested (but please type this underneath the appropriate checkbox).
- This form is based on Algorithms 2 and 3 in the Measure Evaluation Criteria and Guidance document (see pages 18-24). These algorithms provide guidance to help you rate the Reliability and Validity subcriteria. ***We ask that you refer to this document when you are evaluating your measures***.
- Please contact Methods Panel staff if you have questions (methodspanel@qualityforum.org).

Measure Number: 1790
Measure Title: Risk-Adjusted Morbidity and Mortality for Lung Resection for Lung Cancer

RELIABILITY

21. Are submitted specifications precise, unambiguous, and complete so that they can be consistently implemented? *NOTE: NQF staff will conduct a separate, more technical, check of eMeasure (eCQM) specifications, value sets, logic, and feasibility, so no need to consider these in your evaluation.*
    TIPS: Consider the following: Are all the data elements clearly defined? Are all appropriate codes included? Is the logic or calculation algorithm clear? Is it likely this measure can be consistently implemented?
    ☒Yes (go to Question #2)
    ☐No (please explain below, and go to Question #2) NOTE that even though ***non-precise specifications should result in an overall LOW rating for reliability***, we still want you to look at the testing results.

22. Was empirical reliability testing (at the data element or measure score level) conducted using statistical tests with the measure as specified?
    TIPS: Check the 2nd "NO" box below if: only descriptive statistics provided; only describes process for data management/cleaning/computer programming; testing does not match measure specifications (i.e. data, eMeasure, level of analysis, patients)
    ☒Yes (go to Question #4)
    ☐No, there is reliability testing information, but *not* using statistical tests and/or not for the measure as specified OR there is no reliability testing (please explain below then go to

Question #3)

23. Was **empirical VALIDITY testing** of patient-level data conducted?
    - ☐Yes (use your rating from data element validity testing – Question #16- under Validity Section)
    - ☐No (please explain below and rate Question #11: OVERALL RELIABILITY as INSUFFICIENT and proceed to the VALIDITY SECTION)

24. Was reliability testing conducted with computed performance measure scores for each measured entity?
    TIPS: Answer no if: only one overall score for all patients in sample used for testing patient-level data
    - ☒Yes (go to Question #5)
    - ☐No (go to Question #8)

25. Was the method described and appropriate for assessing the proportion of variability due to real differences among measured entities? *NOTE: If multiple methods used, at least one must be appropriate.*
    TIPS: Examples of appropriate methods include signal-to-noise analysis (e.g. Adams/RAND tutorial); random split-half correlation; other accepted method with description of how it assesses reliability of the performance score.
    - ☒Yes (go to Question #6) (The method description could be made better by providing reference(s), slightly more information about Bayesian estimation of the true value. I think the denominator of the equation is missing a superscript. In addition, on page 5, 3rd row from the bottom, one notation is off, theta should be rho.)
    - ☐No (please explain below then go to Question #8)

26. **RATING (score level)** - What is the level of certainty or confidence that the performance measure scores are reliable?
    TIPS: Consider the following: Is the test sample adequate to generalize for widespread implementation? Do the results demonstrate sufficient reliability so that differences in performance can be identified?
    - ☐High (go to Question #8)
    - ☒Moderate (go to Question #8)
    - ☐Low (please explain below then go to Question #7)

27. Was other reliability testing reported?
    - ☐Yes (go to Question #8)
    - ☐No (rate Question #11: OVERALL RELIABILITY as LOW and proceed to the VALIDITY SECTION)

28. Was reliability testing conducted with patient-level data elements that are used to construct the performance measure?
    TIPS: Prior reliability studies of the same data elements may be submitted; if comparing abstraction to "authoritative source/gold standard" see Validity Section Question #15)
    - ☐Yes (go to Question #9)
    - ☒No (if there is score-level testing, rate Question #11: OVERALL RELIABILITY based on score-level rating from Question #6; otherwise, rate Question #11: OVERALL RELIABILITY as INSUFFICIENT. Then proceed to the VALIDITY SECTION)

29. Was the method described and appropriate for assessing the reliability of ALL critical data elements?

TIPS: For example: inter-abstractor agreement (ICC, Kappa); other accepted method with description of how it assesses reliability of the data elements

Answer no if: only assessed percent agreement; did not assess separately for all data elements (at least numerator, denominator, exclusions)

☐Yes (go to Question #10)

☐No (if no, please explain below and rate Question #10 as INSUFFICIENT)

30. **RATING (data element)** – Based on the reliability statistic and scope of testing (number and representativeness of patients and entities), what is the level of certainty or confidence that the data used in the measure are reliable?

TIPS: Consider the following: Is the test sample adequate to generalize for widespread implementation? Can data elements be collected consistently?

☐Moderate (if score-level testing was NOT conducted, rate Question #11: OVERALL RELIABILITY as MODERATE)

☐Low (if score-level testing was NOT conducted, rate Question #11: OVERALL RELIABILITY as LOW)

☐Insufficient (go to Question #11)

11. OVERALL RELIABILITY RATING

OVERALL RATING OF RELIABILITY taking into account precision of specifications and all testing results:

☐High (NOTE: Can be HIGH only if score-level testing has been conducted)

☒Moderate (NOTE: Moderate is the highest eligible rating if score-level testing has not been conducted)

☐Low (please explain below) [NOTE: Should rate LOW if you believe specifications are NOT precise, unambiguous, and complete]

☐Insufficient (please explain below) [NOTE: For most measure types, testing at both the score level and the data element level is not required]

VALIDITY

Assessment of Threats to Validity

33. Were all potential threats to validity that are relevant to the measure empirically assessed?

TIPS: Threats to validity include: exclusions; need for risk adjustment; Able to identify statistically significant and meaningful differences; multiple sets of specifications; missing data/nonresponse.

☒Yes (go to Question #2)

☐No (please explain below and go to Question #2) [NOTE that even if ***non-assessment of applicable threats should result in an overall INSUFFICENT rating for validity***, we still want you to look at the testing results]

34. Analysis of potential threats to validity: Any concerns with measure exclusions?

TIPS: Consider the following: Are the exclusions consistent with the evidence? Are any patients or patient groups inappropriately excluded from the measure? Are the exclusions/exceptions of sufficient frequency and variation across providers to be needed (and outweigh the data collection burden)? If patient preference (e.g., informed decisionmaking) is a basis for exclusion, does it impact performance and if yes, is the measure specified so that the information about patient preference and the effect on the measure is transparent?

☐Yes (please explain below then go to Question #3)

☒No (go to Question #3) (It would be helpful if the developer quantifies the exclusion rates for various exclusion criteria. In the attached paper, it did describe the overall exclusion rate, but it would be better if they provide criterion specific information in the testing form.)

☐Not applicable (i.e., there are no exclusions specified for the measure; go to Question #3)

35. Analysis of potential threats to validity: Risk-adjustment (applies to all outcome, cost, and resource use measures; may also apply to other types of measure)

☐Not applicable (e.g., structure or process measure that is not risk-adjusted; go to Question #4)

a. Is a conceptual rationale for social risk factors included? ☐Yes ☒No (The developer did not conduct any analysis related to social risk factors and did not provide any conceptual rationale for social risk factors. They could consider using other information as a proxy for patient social risk factors, for example, using insurance information to identify dual eligible patients as they mentioned in their testing form.)

b. Are social risk factors included in risk model? ☐Yes ☒No

c. Any concerns regarding the risk-adjustment approach?

*TIPS: Consider the following: If a justification for **not risk adjusting** is provided, is there any evidence that contradicts the developer's rationale and analysis? If the developer asserts there is **no conceptual basis** for adjusting this measure for social risk factors, do you agree with the rationale? **If risk adjusted**: Are the candidate and final variables included in the risk adjustment model adequately described for the measure to be implemented? Are the candidate and final variables included in the risk adjustment model adequately described for the measure to be implemented? Are all of the risk adjustment variables present at the start of care (if not, do you agree with the rationale)? If social risk factors are not included in the risk-adjustment approach, do you agree with the developer's decision? Is an appropriate risk-adjustment strategy included in the measure (e.g., adequate model discrimination and calibration)? Are all statistical model specifications included, including a "clinical model only" if social risk factors are included in the final model?*

☒Yes (please explain below then go to Question #4) (In general, it is fine. But it would be very helpful to have an external validation of the risk adjustment model. That is, the model developed based on the development sample works similarly well in a validation sample.)

☐No (go to Question #4)

36. Analysis of potential threats to validity: Any concerns regarding ability to identify meaningful differences in performance or overall poor performance?

☐Yes (please explain below then go to Question #5)
☒No (go to Question #5)

37. Analysis of potential threats to validity: Any concerns regarding comparability of results if multiple data sources or methods are specified?
☐Yes (please explain below then go to Question #6)
☒No (go to Question #6)
☐Not applicable (go to Question #6)

38. Analysis of potential threats to validity: Any concerns regarding missing data?
☐Yes (please explain below then go to Question #7)
☒No (go to Question #7) (The potential concern is if participating facilities submit all their cases, that is, not selectively submit their cases. The developer mentioned (page 8) "there was consistent agreement across all participants for data completeness." Does this really mean 100% for all sites? If yes, it is good to know; if not, then that needs to be quantified.)

Assessment of Measure Testing

39. Was empirical validity testing conducted using the measure as specified and appropriate statistical test?
Answer no if: face validity; only refer to clinical evidence; only descriptive statistics; only describe process for data management/cleaning/computer programming; testing does not match measure specifications (i.e. data, eMeasure, level, setting, patients).
☒Yes (go to Question #10) [NOTE: If appropriate empirical testing has been conducted, then evaluation of face validity is not necessary. Go to Question #8 **only if** there is insufficient information provided to evaluate data element and score-level testing.]
(Measure testing was done at the hospital level only although the developer checked both hospital and group practice (section 1.4), the developer should uncheck "group practice".).
☐No (please explain below then go to Question #8)

40. Was face validity systematically assessed by recognized experts to determine agreement on whether the computed performance measure score from the measure as specified can be used to distinguish good and poor quality?
TIPS: Answer no if: focused on data element accuracy/availability/feasibility/other topics; the degree of consensus and any areas of disagreement not provided/discussed.
☐Yes (go to Question #9)
☐No (please explain below and rate Question #17: OVERALL VALIDITY as INSUFFICIENT)

41. **RATING (face validity)** - Do the face validity testing results indicate substantial agreement that the performance measure score from the measure as specified can be used to distinguish quality AND potential threats to validity are not a problem, OR are adequately addressed so results are not biased?
☐Yes (if a NEW measure, rate Question #17: OVERALL VALIDITY as MODERATE)
☐ Yes (if a MAINTENANCE measure, do you agree with the justification for not conducting empirical testing? If no, rate Question #17: OVERALL VALIDITY as INSUFFICIENT; otherwise, rate Question #17: OVERALL VALIDITY as MODERATE)
☐No (please explain below and rate Question #17: OVERALL VALIDITY AS LOW)

42. Was validity testing conducted with computed performance measure scores for each measured entity?

TIPS: Answer no if: one overall score for all patients in sample used for testing patient-level data.

☐Yes (go to Question #11)

☒No (please explain below and go to Question #13) (The developer checked "Empirical validity testing (page 7) leaving both critical data elements and performance measure score unchecked.)

43. Was the method described and appropriate for assessing conceptually and theoretically sound hypothesized relationships?

TIPS: For example: correlation of the performance measure score on this measure and other performance measures; differences in performance scores between groups known to differ on quality; other accepted method with description of how it assesses validity of the performance score

☐Yes (go to Question #12)

☐No (please explain below, rate Question #12 as INSUFFICIENT and then go to Question #14)

44. **RATING (measure score)** - Based on the measure score results (significance, strength) and scope of testing (number of measured entities and representativeness) and analysis of potential threats, what is the level of certainty or confidence that the performance measure scores are a valid indicator of quality?

☐High (go to Question #14)

☐Moderate (go to Question #14)

☐Low (please explain below then go to Question #13)

☐Insufficient

45. Was other validity testing reported?

☒Yes (go to Question #14)

☐No (please explain below and rate Question #17: OVERALL VALIDITY as LOW)

46. Was validity testing conducted with patient-level data elements?

TIPS: Prior validity studies of the same data elements may be submitted

☒Yes (go to Question #15)

☐No (please explain below and rate Question #17: OVERALL VALIDITY as INSUFFICIENT if no score-level testing was conducted, otherwise, rate Question #17: OVERALL VALIDITY based on score-level rating from Question #12)

47. Was the method described and appropriate for assessing the accuracy of ALL critical data elements? *NOTE that data element validation from the literature is acceptable.*

TIPS: For example: Data validity/accuracy as compared to authoritative source- sensitivity, specificity, PPV, NPV; other accepted method with description of how it assesses validity of the data elements.

Answer No if: only assessed percent agreement; did not assess separately for all data elements (at least numerator, denominator, exclusions)

☒Yes (go to Question #16)

(A. Concerns for two critical data elements, one is "FEV1 Predicted" (page 8), this is a risk adjustment variable. The agreement rate for this variable is only 76.33%. Another is "Status 30 Days after surgery" (page 9), even though the agreement rate is 92.40%, given that this is a very important endpoint that should have little ambiguity, basically 30-day mortality, I would hope the agreement

rate for this variable is higher. If the agreement varies across hospitals, then it would be even more concerning.

(B. Although the developer did not provide kappa statistic as required due to lack of patient level data, in this case, having percent agreement is sufficiently informative. In fact, in some situations, reporting kappa statistic without percent agreement is worse than reporting percent agreement without kappa. It is known that in some situations the observed proportion of agreement "can be paradoxically altered by the chance-corrected ratio that creates kappa as an index of concordance." (See "High agreement but low kappa: I. The problems of two paradoxes", Feinstein & Cicchetti))

☐No (please explain below and rate Question #16 as INSUFFICIENT)

48. **RATING (data element)** - Based on the data element testing results (significance, strength) and scope of testing (number and representativeness of patients and entities) and analysis of potential threats, what is the level of certainty or confidence that the data used in the measure are valid?

☒Moderate (if score-level testing was NOT conducted, rate Question #17: OVERALL VALIDITY as MODERATE) (I don't see score level validity testing, some concerns about two critical data elements.)

☐Low (please explain below) (if score-level testing was NOT conducted, rate Question #17: OVERALL VALIDITY as LOW)

☐Insufficient (go to Question #17)

17. OVERALL VALIDITY RATING

OVERALL RATING OF VALIDITY taking into account the results and scope of all testing and analysis of potential threats.

☐High (NOTE: Can be HIGH only if score-level testing has been conducted)

☒Moderate (NOTE: Moderate is the highest eligible rating if score-level testing has NOT been conducted)

☐Low (please explain below) [NOTE: Should rate LOW if you believe that there are threats to validity and/or threats to validity were not assessed]

☐Insufficient (if insufficient, please explain below) [NOTE: For most measure types, testing at both the score level and the data element level is not required] [NOTE: If rating is INSUFFICIENT for all empirical testing, then go back to Question #8 and evaluate any face validity that was conducted, then reconsider this overall rating.]

FOR COMPOSITE MEASURES ONLY: Empirical analyses to support composite construction

What is the level of certainty or confidence that the empirical analysis demonstrates that the component measures add value to the composite and that the aggregation and weighting rules are consistent with the quality construct?

TIPS: Consider the following: Do the component measures fit the quality construct? Are the objectives of parsimony and simplicity achieved while supporting the quality construct?

☐High

☐Moderate

☐Low (please explain below)

☐Insufficient (please explain below)

NATIONAL QUALITY FORUM—Evidence (subcriterion 1a)

Measure Number (*if previously endorsed*)**:** 1790

Measure Title: Risk-Adjusted Morbidity and Mortality for Lung Resection for Lung Cancer

IF the measure is a component in a composite performance measure, provide the title of the Composite Measure here:

Date of Submission: 11/15/2017

Instructions

- *Complete 1a.1 and 1a.2 for all measures. If instrument-based measure, complete 1a.3.*
- *Complete **EITHER 1a.2, 1a.3 or 1a.4** as applicable for the type of measure and evidence.*
- *For composite performance measures:*
  - *A separate evidence form is required for each component measure unless several components were studied together.*
  - *If a component measure is submitted as an individual performance measure, attach the evidence form to the individual measure submission.*
- All information needed to demonstrate meeting the evidence subcriterion (1a) must be in this form. An appendix of *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.

Note: The information provided in this form is intended to aid the Standing Committee and other stakeholders in understanding to what degree the evidence for this measure meets NQF's evaluation criteria.

1a. Evidence to Support the Measure Focus

The measure focus is evidence-based, demonstrated as follows:

- Outcome: [3] Empirical data demonstrate a relationship between the outcome and at least one healthcare structure, process, intervention, or service. If not available, wide variation in performance can be used as evidence, assuming the data are from a robust number of providers and results are not subject to systematic bias.
- Intermediate clinical outcome: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured intermediate clinical outcome leads to a desired health outcome.
- Process: [5] a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured process leads to a desired health outcome.
- Structure: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured structure leads to a desired health outcome.
- Efficiency: [6] evidence not required for the resource use component.
- For measures derived from patient reports, evidence should demonstrate that the target population values the measured outcome, process, or structure and finds it meaningful.
- Process measures incorporating Appropriate Use Criteria: See NQF's guidance for evidence for measures, in general; guidance for measures specifically based on clinical practice guidelines apply as well.

Notes

3. Generally, rare event outcomes do not provide adequate information for improvement or discrimination; however, serious reportable events that are compared to zero are appropriate outcomes for public reporting and quality improvement.

4. The preferred systems for grading the evidence are the Grading of Recommendations, Assessment, Development and Evaluation (GRADE) guidelines and/or modified GRADE.

5. Clinical care processes typically include multiple steps: assess → identify problem/potential problem → choose/plan intervention (with patient input) → provide intervention → evaluate impact on health status. If the measure focus is one step in such a multistep process, the step with the

strongest evidence for the link to the desired outcome should be selected as the focus of measurement. Note: A measure focused only on collecting PROM data is not a PRO-PM.

6. Measures of efficiency combine the concepts of resource use and quality (see NQF's Measurement Framework: Evaluating Efficiency Across Episodes of Care; AQA Principles of Efficiency Measures).

1a.1.This is a measure of: (*should be consistent with type of measure entered in De.1*)

Outcome

☒ Outcome: Postoperative complications: reintubation, need for tracheostomy, initial ventilator support > 48 hours, ARDS, pneumonia, pulmonary embolus, bronchopleural fistula, unexpected return to the operating room, myocardial infarction or operative mortality (death during the index hospitalization, regardless of timing, or within 30 days, regardless of location).

☐Patient-reported outcome (PRO):

PROs include HRQoL/functional status, symptom/symptom burden, experience with care, health-related behaviors. (A PRO-based performance measure is not a survey instrument. Data may be collected using a survey instrument to construct a PRO measure.)

☐ Intermediate clinical outcome (*e.g., lab value*):

☐ Process:

☐ Appropriate use measure:

☐ Structure:

☐ Composite:

1a.2 LOGIC MODEL Diagram or briefly describe the steps between the healthcare structures and processes (e.g., interventions, or services) and the patient's health outcome(s). The relationships in the diagram should be easily understood by general, non-technical audiences. Indicate the structure, process or outcome being measured.

Postoperative complications and operative mortality are important negative outcomes associated with lung cancer resection surgery. The STS lung cancer resection risk model (Fernandez et al, 2016) identifies predictors of these outcomes, including patient age, smoking status, comorbid medical conditions, and other patient characteristics, as well as operative approach and the extent of pulmonary resection. Knowledge of these predictors informs clinical decision making by enabling physicians and patients to understand the associations between individual patient characteristics and outcomes and – with continuous feedback of performance data over time – fosters quality improvement.

Fernandez FG, Kosinski AS, Burfeind W, et al. The Society of Thoracic Surgeons lung cancer resection risk model: higher quality data and superior outcomes. Ann Thorac Surg 2016;102:370-7.

1a.3 Value and Meaningfulness: IF this measure is derived from patient report, provide evidence that the target population values the measured ***outcome, process, or structure*** and finds it meaningful. (Describe how and from whom their input was obtained.)

n/a

****RESPOND TO ONLY ONE SECTION BELOW -EITHER 1a.2, 1a.3 or 1a.4) ****

1a.2 FOR OUTCOME MEASURES including PATIENT REPORTED OUTCOMES - Provide empirical data demonstrating the relationship between the outcome (or PRO) to at least one healthcare structure, process, intervention, or service.

The STS lung cancer resection data demonstrate a significant relationship between operative approach (i.e., thoracoscopy vs. thoracotomy), postoperative complications and operative mortality. Please see Table 4 in the attachment (Fernandez et al, 2016) for empirical data related to operative approach and also for procedure type/extent of pulmonary resection.

Fernandez FG, Kosinski AS, Burfeind W, et al. The Society of Thoracic Surgeons lung cancer resection risk model: higher quality data and superior outcomes. Ann Thorac Surg 2016;102:370-7.

1a.3. SYSTEMATIC REVIEW(SR) OF THE EVIDENCE (for INTERMEDIATE OUTCOME, PROCESS, OR STRUCTURE PERFORMANCE MEASURES, INCLUDING THOSE THAT ARE INSTRUMENT-BASED) If the evidence is not based on a systematic review go to section 1a.4) If you wish to include more than one systematic review, add additional tables.

What is the source of the systematic review of the body of evidence that supports the performance measure? A systematic review is a scientific investigation that focuses on a specific question and uses explicit, prespecified scientific methods to identify, select, assess, and summarize the findings of similar but separate studies. It may include a quantitative synthesis (meta-analysis), depending on the available data. (IOM)

☐Clinical Practice Guideline recommendation (with evidence review)

☐US Preventive Services Task Force Recommendation

☐Other systematic review and grading of the body of evidence (*e.g., Cochrane Collaboration, AHRQ Evidence Practice Center*)

☐Other

Source of Systematic Review:	

• **Title** • **Author** • **Date** • **Citation, including page number** • **URL**	
Quote the guideline or recommendation verbatim about the process, structure or intermediate outcome being measured. If not a guideline, summarize the conclusions from the SR.	
Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	
Provide all other grades and definitions from the evidence grading system	
Grade assigned to the **recommendation** with definition of the grade	
Provide all other grades and definitions from the recommendation grading system	
Body of evidence: • Quantity – how many studies? • Quality – what type of studies?	
Estimates of benefit and consistency across studies	
What harms were identified?	
Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	

1a.4 OTHER SOURCE OF EVIDENCE

If source of evidence is NOT from a clinical practice guideline, USPSTF, or systematic review, please describe the evidence on which you are basing the performance measure.

1a.4.1 Briefly SYNTHESIZE the evidence that supports the measure. A list of references without a summary is not acceptable.

1a.4.2 What process was used to identify the evidence?

1a.4.3. Provide the citation(s) for the evidence.

1. Evidence, Performance Gap, Priority – Importance to Measure and Report
Extent to which the specific measure focus is evidence-based, important to making significant gains in healthcare quality, and improving health outcomes for a specific high-priority (high-impact) aspect of healthcare where there is variation in or overall less-than-optimal performance. ***Measures must be judged to meet all sub criteria to pass this criterion and be evaluated against the remaining criteria.***
1a. Evidence to Support the Measure Focus – See attached Evidence Submission Form NQF_evidence_attachment_STS-1790-111517-v2.docx **1a.1 <u>For Maintenance of Endorsement:</u> Is there new evidence about the measure since the last update/submission?** Do not remove any existing information. If there have been any changes to evidence, the Committee will consider the new evidence. Please use the most current version of the evidence attachment (v7.1). Please use red font to indicate updated evidence. No
1b. Performance Gap Demonstration of quality problems and opportunity for improvement, i.e., data demonstrating: • considerable variation, or overall less-than-optimal performance, in the quality of care across providers; and/or • Disparities in care across population groups. **1b.1. Briefly explain the rationale for this measure** *(e.g., how the measure will improve the quality of care, the benefits or improvements in quality envisioned by use of this measure)* *<u>If a COMPOSITE</u> (e.g., combination of component measure scores, all-or-none, any-or-none), SKIP this question and answer the composite questions.*

Providing outcomes data to participating thoracic surgery sites allows benchmarking of practice group results against the STS national results and allows demonstration of improvement when QI efforts are undertaken. These outcomes data aid clinicians and patients in making informed clinical decisions and also enable them to compare risk-adjusted outcomes for quality improvement purposes.

1b.2. **Provide performance scores on the measure as specified (current and over time) at the specified level of analysis**. *(This is required for maintenance of endorsement. Include mean, std dev, min, max, interquartile range, scores by decile. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities include.) This information also will be used to address the sub-criterion on improvement (4b1) under Usability and Use.*
The endpoint of mortality or major morbidity occurred in 9.5% of eligible patients. There is no overlap in credible intervals for hospital-specific SIR between some of the best performing sites (3.5%; 8 of 231 sites with upper limit below 1) and worst performing sites (6.9%; 16 of 231 sites with lower limit above 1), indicating that this model provides meaningful discrimination between best and worst performers.

Dates: January 1, 2012 through December 31, 2014

Data/Sample: The population included 27,844 records from 231 hospitals. Hospital-specific sample sizes ranged from 1 to 852 records per hospital (mean=121, median=85, IQR=[36, 165]).

Distribution of hospital-specific estimates of standardized incidence ratio (SIR) for composite of mortality and morbidity:

Minimum	0.47
1st quartile	0.90
Median	1.00
Mean	1.05
3rd quartile	1.22
Maximum	2.37

1b.3. **If no or limited performance data on the measure as specified is reported in 1b2, then provide a summary of data from the literature that indicates opportunity for improvement or overall less than optimal performance on the specific focus of measurement.**
n/a (see data reported in 1b2)

1b.4. **Provide disparities data from the measure as specified (current and over time) by population group, e.g., by race/ethnicity, gender, age, insurance status, socioeconomic status, and/or disability.** *(This is required for maintenance of endorsement. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included.) For measures that show high levels of performance, i.e., "topped out", disparities data may demonstrate an opportunity for improvement/gap in care for certain sub-populations. This information also will be used to address the sub-criterion on improvement (4b1) under Usability and Use.*
Data/Sample: The population included 27,844 records from 231 hospitals.

Dates: January 1, 2012 through December 31, 2014

Race: White 24,099; Black 2,369; Other 1,217

Incidence of mortality or major morbidity endpoints:
White: 9.8%, 95% CI [9.4,% 10.1%]
Black: 8.9%, 95% CI [7.8%, 10.1%]
Other: 6.9%, 95% CI [5.6, 8.5%]

1b.5. **If no or limited data on disparities from the measure as specified is reported in 1b.4, then provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement. Include citations. Not necessary if performance data provided in 1b.4**
n/a (see data reported in 1b4)

2. Reliability and Validity—Scientific Acceptability of Measure Properties

Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. ***Measures must be judged to meet the sub criteria for both reliability and validity to pass this criterion and be evaluated against the remaining criteria.***

2a.1. Specifications The measure is well defined and precisely specified so it can be implemented consistently within and across organizations and allows for comparability. eMeasures should be specified in the Health Quality Measures Format (HQMF) and the Quality Data Model (QDM).

De.5. **Subject/Topic Area** *(check all the areas that apply):*
Cancer, Cancer : Lung, Esophageal, Surgery, Surgery : Thoracic Surgery

De.6. **Non-Condition Specific***(check all the areas that apply):*
Safety, Safety : Complications

De.7. **Target Population Category** *(Check all the populations for which the measure is specified and tested if any):*
Elderly

S.1. **Measure-specific Web Page** *(Provide a URL link to a web page specific for this measure that contains current detailed specifications including code lists, risk model details, and supplemental materials. Do not enter a URL linking to a home page or to general information.)*
http://www.sts.org/sites/default/files/documents/STSThoracicDataSpecsV2_3.pdf

S.2a. **If this is an eMeasure**, HQMF specifications must be attached. Attach the zipped output from the eMeasure authoring tool (MAT) - if the MAT was not used, contact staff. (Use the specification fields in this online form for the plain-language description of the specifications)
This is not an eMeasure **Attachment:**

S.2b. **Data Dictionary, Code Table, or Value Sets** *(and risk model codes and coefficients when applicable) must be attached. (Excel or csv file in the suggested format preferred - if not, contact staff)*
Attachment **Attachment:** STSThoracicDataSpecsV2_3.pdf

S.2c. Is this an instrument-based measure (i.e., data collected via instruments, surveys, tools, questionnaires, scales, etc.)? Attach copy of instrument if available.
No, this is not an instrument-based measure **Attachment:**

S.2d. Is this an instrument-based measure (i.e., data collected via instruments, surveys, tools, questionnaires, scales, etc.)? Attach copy of instrument if available.
Not an instrument-based measure

S.3.1. **For maintenance of endorsement:** Are there changes to the specifications since the last updates/submission. If yes, update the specifications for S1-2 and S4-22 and explain reasons for the changes in S3.2.
Yes

S.3.2. **For maintenance of endorsement,** please briefly describe any important changes to the measure specifications since last measure update and explain the reasons.
Among postoperative complications included in the numerator statement, "bleeding requiring reoperation" was replaced by "unexpected return to the operating room." Bleeding is only one of many possible reasons for a reoperation; other reasons may include prolonged air leak and chylothorax. STS General Thoracic surgeon leaders felt that the new, expanded definition of reoperation ("unexpected return to the operating room") better reflects the scope of this category of postoperative complications.

S.4. **Numerator Statement** *(Brief, narrative description of the measure focus or what is being measured about the target population, i.e., cases from the target population with the target process, condition, event, or outcome) DO NOT include the rationale for the measure.*

IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm (S.14).
Number of patients greater than or equal to 18 years of age undergoing elective lung resection (Open or VATS wedge resection, segmentectomy, lobectomy, bilobectomy, sleeve lobectomy, pneumonectomy) for lung cancer who developed any of the following postoperative complications: reintubation, need for tracheostomy, initial ventilator support > 48 hours, ARDS, pneumonia, pulmonary embolus, bronchopleural fistula, unexpected return to the operating room, myocardial infarction or operative mortality (death during the index hospitalization, regardless of timing, or within 30 days, regardless of location).

S.5. **Numerator Details** *(All information required to identify and calculate the cases from the target population with the target process, condition, event, or outcome such as definitions, time period for data collection, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
IF an OUTCOME MEASURE, describe how the observed outcome is identified/counted. Calculation of the risk-adjusted outcome should be described in the calculation algorithm (S.14).
Number of patients undergoing elective lung resection for lung cancer for whom:

1. Postoperative events (POEvents - STS GTS Database, v 2.2, sequence number 1710) is marked "Yes" and one of the following items is marked:
a. Reintubation (Reintube - STS GTS Database, v 2.2, sequence number 1850)
b. Need for tracheostomy (Trach - STS GTS Database, v 2.2, sequence number 1860)
c. Initial ventilator support > 48 hours (Vent- STS GTS Database, v 2.2, sequence number 1840)
d. Acute Respiratory Distress Syndrome (ARDS - STS GTS Database, v 2.2, sequence number 1790)
e. Pneumonia (Pneumonia - STS GTS Database, v 2.2, sequence number 1780)
f. Pulmonary Embolus (PE - STS GTS Database, v 2.2, sequence number 1820)
g. Bronchopleural Fistula (Bronchopleural - STS GTS Database, v 2.2, sequence number 1810)
h. Myocardial infarction (MI - STS GTS Database, v 2.2, sequence number 1900)

Or

2. Unexpected return to the operating room (ReturnOR - STS GTS Database, Version 2.2, sequence number 1720) is marked "yes"

Or

3. One of the following fields is marked "dead"
a. Discharge status (MtDCStat - STS GTS Database, Version 2.2, sequence number 2200);
b. Status at 30 days after surgery (Mt30Stat - STS GTS Database, Version 2.2, sequence number 2240)

Please see STS General Thoracic Surgery Database Data Collection Form, Version 2.3-
http://www.sts.org/sites/default/files/documents/STSThoracicDCF_V2_3_MajorProc_Annotated.pdf

S.6. **Denominator Statement** *(Brief, narrative description of the target population being measured)*
Number of patients greater than or equal to 18 years of age undergoing elective lung resection (Open or VATS wedge resection, segmentectomy, lobectomy, bilobectomy, sleeve lobectomy, pneumonectomy) for lung cancer

S.7. **Denominator Details** *(All information required to identify and calculate the target population/denominator such as definitions, time period for data collection, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b.)*
IF an OUTCOME MEASURE, describe how the target population is identified. Calculation of the risk-adjusted outcome should be described in the calculation algorithm (S.14).
1. Lung cancer (LungCancer - STS GTS Database, v 2.2, sequence number 830) is marked "yes" and Category of Disease – Primary (CategoryPrim - STS GTS Database, v 2.2, sequence number 1300) is marked as one of the following:

(ICD-9, ICD-10)
Lung cancer, main bronchus, carina (162.2, C34.00)
Lung cancer, upper lobe (162.3, C34.10)

Lung cancer, middle lobe (162.4, C34.2)
Lung cancer, lower lobe (162.5, C34.30)
Lung cancer, location unspecified (162.9, C34.90)

2. Patient has lung cancer (as defined in #1 above) and primary procedure is one of the following CPT codes:

Thoracoscopy, surgical; with lobectomy (32663)
Thoracoscopy with therapeutic wedge resection (eg mass or nodule) initial, unilateral (32666)
Thoracoscopy with removal of a single lung segment (segmentectomy) (32669)
Thoracoscopy with removal of two lobes (bilobectomy) (32670)
Thoracoscopy with removal of lung, pneumonectomy (32671)
Thoracotomy with therapeutic wedge resection (eg mass nodule) initial (32505)
Removal of lung, total pneumonectomy; (32440)
Removal of lung, single lobe (lobectomy) (32480)
Removal of lung, two lobes (bilobectomy) (32482)
Removal of lung, single segment (segmentectomy) (32484)
Removal of lung, sleeve lobectomy (32486)

3. Status of Operation (Status - STS General Thoracic Surgery Database, Version 2.2, sequence number 1420) is marked as "Elective"

4. Only analyze the first operation of the hospitalization meeting criteria 1-3

S.8. **Denominator Exclusions** *(Brief narrative description of exclusions from the target population)*
Patients were excluded if they had an extrapleural pneumonectomy, completion pneumonectomy, carinal pneumonectomy, occult carcinoma or benign disease on final pathology, or an urgent, emergent, or palliative operation. Furthermore, patients with missing age, sex, discharge mortality status, and predicted forced expiratory volume in 1 second were also excluded.

S.9. **Denominator Exclusion Details** *(All information required to identify and calculate exclusions from the denominator such as definitions, time period for data collection, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b.)*
Cases removed from calculations if any of following fields are checked on the data collection form:

Removal of lung, sleeve (carinal) pneumonectomy (32442)
Removal of lung, total pneumonectomy; extrapleural (32445)
Removal of lung, completion pneumonectomy (32488)

OR if either of the following fields are checked:

Carcinoid tumor of bronchus and lung; benign, typical (209.61., D34.090)
Lung tumor, benign (212.3, D14.30)

OR if Emergent, Urgent, or Palliative is checked under "Status of Operation"

Only general thoracic procedures coded as primary lung or primary esophageal cancer are included in measure calculations, so occult carcinoma is effectively excluded.

S.10. **Stratification Information** *(Provide all information required to stratify the measure results, if necessary, including the stratification variables, definitions, specific data collection items/responses, code/value sets, and the risk-model covariates and coefficients for the clinically-adjusted version of the measure when appropriate – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format with at S.2b.)*
n/a

S.11. **Risk Adjustment Type** (Select type. Provide specifications for risk stratification in measure testing attachment)
Statistical risk model
If other:

S.12. **Type of score:**

Rate/proportion
If other:

S.13. **Interpretation of Score** *(Classifies interpretation of score according to whether better quality is associated with a higher score, a lower score, a score falling within a defined interval, or a passing score)*
Better quality = Lower score

S.14. **Calculation Algorithm/Measure Logic** (*Diagram or describe the calculation of the measure score as an ordered sequence of steps including identifying the target population; exclusions; cases meeting the target process, condition, event, or outcome; time period for data, aggregating data; risk adjustment; etc.)*
Target population is patients undergoing elective lung resection for lung cancer. Emergency procedures were excluded. Outcome is operative mortality (death during the index hospitalization, regardless of timing, or within 30 days, regardless of location) or occurrence of any of the following postoperative complications: reintubation, need for tracheostomy, initial ventilator support > 48 hours, ARDS, pneumonia, pulmonary embolus, bronchopleural fistula, unexpected return to the operating room, or myocardial infarction. Analysis considered 27,844 patients with procedures between 01/01/2012 and 12/31/2014 (36 months). Risk adjustment was achieved with a Bayesian hierarchical model with composite of the above postoperative complications as the outcome. The measure score was estimated with this model.

For additional information, please review the risk model in the attachment. (Fernandez, et. al. 2016.)

S.15. **Sampling** *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*
IF an instrument-based performance measure (e.g., PRO-PM), identify whether (and how) proxy responses are allowed.
n/a

S.16. **Survey/Patient-reported data** *(If measure is based on a survey or instrument, provide instructions for data collection and guidance on minimum response rate.)*
Specify calculation of response rates to be reported with performance measure results.
n/a

S.17. **Data Source** *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*
If other, please describe in S.18.
Other, Registry Data

S.18. **Data Source or Collection Instrument** *(Identify the specific data source/data collection instrument (e.g. name of database, clinical registry, collection instrument, etc., and describe how data are collected.)*
IF instrument-based, identify the specific instrument(s) and standard methods, modes, and languages of administration.
STS General Thoracic Surgery Database, Version 2.3

S.19. **Data Source or Collection Instrument** *(available at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
Available at measure-specific web page URL identified in S.1

S.20. **Level of Analysis** *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*
Facility

S.21. **Care Setting** *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*
Inpatient/Hospital
If other:

S.22. **COMPOSITE Performance Measure** - Additional Specifications *(Use this section as needed for aggregation and weighting rules, or calculation of individual performance measures if not individually endorsed.)*

2. Validity – See attached Measure Testing Submission Form

2.1 For maintenance of endorsement

Reliability testing: If testing of reliability of the measure score was not presented in prior submission(s), has reliability testing of the measure score been conducted? If yes, please provide results in the Testing attachment. Please use the most current version of the testing attachment (v7.1). Include information on all testing conducted (prior testing as well as any new testing); use red font to indicate updated testing.
Yes

2.2 For maintenance of endorsement
Has additional empirical validity testing of the measure score been conducted? If yes, please provide results in the Testing attachment. Please use the most current version of the testing attachment (v7.1). Include information on all testing conducted (prior testing as well as any new testing); use red font to indicate updated testing.
Yes

2.3 For maintenance of endorsement
Risk adjustment: For outcome, resource use, cost, and some process measures, risk-adjustment that includes social risk factors is not prohibited at present. Please update sections 1.8, 2a2, 2b1,2b4.3 and 2b5 in the Testing attachment and S.140 and S.11 in the online submission form. NOTE: These sections must be updated even if social risk factors are not included in the risk-adjustment strategy. You MUST use the most current version of the Testing Attachment (v7.1) -- older versions of the form will not have all required questions.
Yes - Updated information is included

NATIONAL QUALITY FORUM—Measure Testing (subcriteria 2a2, 2b1-2b6)

Measure Number (*if previously endorsed*)**:** 1790
Measure Title: Risk-Adjusted Morbidity and Mortality for Lung Resection for Lung Cancer
Date of Submission: 11/15/2017
Type of Measure:

☒ Outcome (*including PRO-PM*)	☐ Composite – ***STOP – use composite testing form***
☐Intermediate Clinical Outcome	☐ Cost/resource
☐ Process *(including Appropriate Use)*	☐ Efficiency
☐ Structure	

Instructions

- Measures must be tested for all the data sources and levels of analyses that are specified. ***If there is more than one set of data specifications or more than one level of analysis, contact NQF staff*** about how to present all the testing information in one form.
- **For all measures, sections 1, 2a2, 2b1, 2b2, and 2b4 must be completed.**
- **For outcome and resource use measures**, section **2b3** also must be completed.
- If specified for **multiple data sources/sets of specificaitons** (e.g., claims and EHRs), section **2b5** also must be completed.
- Respond to all questions as instructed with answers immediately following the question. All information on testing to demonstrate meeting the subcriteria for reliability (2a2) and validity (2b1-2b6) must be in this form. An appendix for *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.

- Maximum of 25 pages (*incuding questions/instructions;* minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.
- For information on the most updated guidance on how to address social risk factors variables and testing in this form refer to the release notes for version 7.1 of the Measure Testing Attachment.

Note: The information provided in this form is intended to aid the Standing Committee and other stakeholders in understanding to what degree the testing results for this measure meet NQF's evaluation criteria for testing.

2a2. Reliability testing [10] demonstrates the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise. For **instrument-based measures** (including PRO-PMs) **and composite performance measures**, reliability should be demonstrated for the computed performance score.

2b1. Validity testing [11] demonstrates that the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality. For **instrument-based measures (including PRO-PMs) and composite performance measures**, validity should be demonstrated for the computed performance score.

2b2. Exclusions are supported by the clinical evidence and are of sufficient frequency to warrant inclusion in the specifications of the measure; [12]

AND

If patient preference (e.g., informed decisionmaking) is a basis for exclusion, there must be evidence that the exclusion impacts performance on the measure; in such cases, the measure must be specified so that the information about patient preference and the effect on the measure is transparent (e.g., numerator category computed separately, denominator exclusion category computed separately). [13]

2b3. For outcome measures and other measures when indicated (e.g., resource use):

- **an evidence-based risk-adjustment strategy** (e.g., risk models, risk stratification) is specified; is based on patient factors (including clinical and social risk factors) that influence the measured outcome and are present at start of care; [14,15] and has demonstrated adequate discrimination and calibration

OR

- rationale/data support no risk adjustment/ stratification.

2b4. Data analysis of computed measure scores demonstrates that methods for scoring and analysis of the specified measure allow for **identification of statistically significant and practically/clinically meaningful** [16] **differences in performance**;

OR

there is evidence of overall less-than-optimal performance.

2b5. If multiple data sources/methods are specified, there is demonstration they produce comparable results.

2b6. Analyses identify the extent and distribution of **missing data** (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias.

Notes

10. Reliability testing applies to both the data elements and computed measure score. Examples of reliability testing for data elements include, but are not limited to: inter-rater/abstractor or intra-rater/abstractor studies; internal consistency for multi-item scales; test-retest for survey items. Reliability testing of the measure score addresses precision of measurement (e.g., signal-to-noise).

11. Validity testing applies to both the data elements and computed measure score. Validity testing of data elements typically analyzes agreement with another authoritative source of the same information. Examples of validity testing of the measure score include, but are not limited to: testing hypotheses that the measures scores indicate quality of care, e.g., measure scores are different for groups known to have differences in quality assessed by another valid quality measure or method; correlation of measure scores with another valid indicator of quality for the specific topic; or relationship to conceptually related measures (e.g., scores on process measures to scores on outcome measures). Face validity of the measure score as a quality indicator may be adequate if accomplished through a systematic and transparent process, by identified experts, and explicitly addresses whether performance scores resulting from the measure as specified can be used to distinguish good from poor quality. The degree of consensus and any areas of disagreement must be provided/discussed.

12. Examples of evidence that an exclusion distorts measure results include, but are not limited to: frequency of occurrence, variability of exclusions across providers, and sensitivity analyses with and without the exclusion.

13. Patient preference is not a clinical exception to eligibility and can be influenced by provider interventions.

14. Risk factors that influence outcomes should not be specified as exclusions.

15. With large enough sample sizes, small differences that are statistically significant may or may not be practically or clinically meaningful. The substantive question may be, for example, whether a statistically significant difference of one percentage point in the percentage of patients who received smoking cessation counseling (e.g., 74 percent v. 75 percent) is clinically meaningful; or whether a statistically significant difference of $25 in cost for an episode of care (e.g., $5,000 v. $5,025) is practically meaningful. Measures with overall less-than-optimal performance may not demonstrate much variability across providers.

1. DATA/SAMPLE USED FOR ALL TESTING OF THIS MEASURE

Often the same data are used for all aspects of measure testing. In an effort to eliminate duplication, the first five questions apply to all measure testing. If there are differences by aspect of testing,(e.g., reliability vs. validity) be sure to indicate the specific differences in question 1.7.

1.1. What type of data was used for testing? (*Check all the sources of data identified in the measure specifications and data used for testing the measure. Testing must be provided for all the sources of data specified and intended for measure implementation.* ***If different data sources are used for the numerator and denominator, indicate N [numerator] or D [denominator] after the checkbox.***)

Measure Specified to Use Data From: (***must be consistent with data sources entered in S.17***)	**Measure Tested with Data From:**
☐ abstracted from paper record	☐ abstracted from paper record
☐ claims	☐ claims
☒ registry	☒ registry
☐ abstracted from electronic health record	☐ abstracted from electronic health record
☐ eMeasure (HQMF) implemented in EHRs	☐ eMeasure (HQMF) implemented in EHRs
☐ other:	☐ other:

1.2. If an existing dataset was used, identify the specific dataset (*the dataset used for testing must be consistent with the measure specifications for target population and healthcare entities being measured; e.g.,*

Medicare Part A claims, Medicaid claims, other commercial insurance, nursing home MDS, home health OASIS, clinical registry).
STS General Thoracic Surgery Database, Version 2.2

1.3. What are the dates of the data used in testing? 01/01/2012 – 12/31/2014

1.4. What levels of analysis were tested? (*testing must be provided for <u>all</u> the levels specified and intended for measure implementation, e.g., individual clinician, hospital, health plan*)

Measure Specified to Measure Performance of: (***must be consistent with levels entered in item S.20***)	**Measure Tested at Level of:**
☐ individual clinician	☐ individual clinician
☒ group/practice	☒ group/practice
☒ hospital/facility/agency	☒ hospital/facility/agency
☐ health plan	☐ health plan
☐ other:	☐ other:

1.5. How many and which <u>measured entities</u> were included in the testing and analysis (by level of analysis and data source)? (*identify the number and descriptive characteristics of measured entities included in the analysis (e.g., size, location, type); if a sample was used, describe how entities were selected for inclusion in the sample*)
The analysis population consisted of all STS records for patients meeting measure inclusion criteria who had their surgery during January 1, 2012 through December 31, 2014. The population included 27,844 records from 231 hospitals. Hospital-specific sample sizes ranged from 1 to 852 records per hospital (mean=121, median=85, IQR=[36, 165]).

1.6. How many and which <u>patients</u> were included in the testing and analysis (by level of analysis and data source)? (*identify the number and descriptive characteristics of patients included in the analysis (e.g., age, sex, race, diagnosis); if a sample was used, describe how patients were selected for inclusion in the sample*)

Patient Characteristics [n (%) or mean ± SD].

Variable	Values
Total	27,844 (100)
Age, years	67.2 ± 10.1
Male	12,647 (45.4)
Race	
White	24,099 (87.0)
Black	2,369 (8.6)
Other	1,217 (4.4)
Body mass index, kg/m^2[a]	27.6 ± 6.2
Coronary artery disease	6,196 (22.3)
Diabetes mellitus	5,158 (18.5)
Renal dysfunction	504 (1.8)
Induction chemotherapy or radiation	1,801 (6.5)
Cigarette smoking	
Never	3,895 (14.0)
Past (stopped more than 1 month)	17,368 (62.4)
Current	6,581 (23.6)
Steroids	965 (3.5)
Minimally invasive	17,153 (61.6)
Thoracotomy	10,691 (38.4)
Primary procedure	
Wedge resection	3,815 (13.7)
Segmentectomy	1,685 (6.1)
Lobectomy	19,836 (71.2)
Sleeve lobectomy	412 (1.5)
Bilobectomy	980 (3.5)
Pneumonectomy	1,116 (4.0)

1.7. If there are differences in the data or sample used for different aspects of testing (e.g., reliability, validity, exclusions, risk adjustment), identify how the data or sample are different for each aspect of testing reported below.

The STS tests reliability based on three years of data in the General Thoracic Surgery Database (see 1.5 above). Validity testing is conducted on an annual basis through the audit of data completeness and accuracy in randomly-selected surgical records at randomly-selected GTSD participant sites (see 2b1.2 below).

1.8 What were the social risk factors that were available and analyzed? For example, patient-reported data (e.g., income, education, language), proxy variables when social risk data are not collected from each patient (e.g. census tract), or patient community characteristics (e.g. percent vacant housing, crime rate) which do not have to be a proxy for patient-level data.

Patient social risk data are not collected in the General Thoracic Surgery Database. Through the collection of insurance information, information on dual Medicare/Medicaid eligibility is available from the database, which can serve as a proxy for low income and patient vulnerability. However, this information is not presently included in STS data analysis nor as a basis for stratification in STS measures.

2a2. RELIABILITY TESTING

***Note**: If accuracy/correctness (validity) of data elements was empirically tested, separate reliability testing of data elements is not required – in 2a2.1 check critical data elements; in 2a2.2 enter "see section 2b2 for validity testing of data elements"; and skip 2a2.3 and 2a2.4.*

2a2.1. What level of reliability testing was conducted? (*may be one or both levels*)
☐ **Critical data elements used in the measure** (*e.g., inter-abstractor reliability; data element reliability must address ALL critical data elements*)
☒ **Performance measure score** (e.g., *signal-to-noise analysis*)

2a2.2. For each level checked above, describe the method of reliability testing and what it tests (*describe the steps—do not just name a method; what type of error does it test; what statistical analysis was used*)

Reliability is conventionally defined as the proportion of variation in a performance measure that is due to true between-hospital differences (i.e., signal) as opposed to random statistical fluctuations (i.e., noise). A mathematically equivalent definition is the squared correlation between a measurement and the true value. We estimated this quantity within the Bayesian statistical framework. We computed the squared correlation between each hospital's estimated performance measure (the estimated SIR) and the true value (estimated using Bayesian inference methods). Accordingly, reliability was defined as the square of the Pearson correlation coefficient (ρ^2) between the set of participant-specific estimates $\hat{\theta}_1, \ldots, \hat{\theta}_N$ and the corresponding unknown true values, $\theta_1,\ldots,\theta_N$, that is:

$$\rho^2 = \frac{\sum_{j=1}^{N}(\hat{\theta}_j - \frac{1}{N}\sum_{h=1}^{N}\hat{\theta}_h)(\theta_j - \frac{1}{N}\sum_{h=1}^{N}\theta_h)}{\sum_{j=1}^{N}(\hat{\theta}_j - \frac{1}{N}\sum_{h=1}^{N}\hat{\theta}_h)^2 \sum_{j=1}^{N}(\theta_j - \frac{1}{N}\sum_{h=1}^{N}\theta_h)^2}$$

The quantity ρ^2 was estimated by its posterior mean, namely,

$$\hat{\rho}^2 = \frac{1}{5000}\sum_{l=1}^{5000}\rho^2_{(l)}$$

where

$$\rho^2_{(l)} = \frac{\sum_{j=1}^{N}(\hat{\theta}_j - \frac{1}{N}\sum_{h=1}^{N}\hat{\theta}_h)(\theta_j^{(l)} - \frac{1}{N}\sum_{h=1}^{N}\theta_h^{(l)})}{\sum_{j=1}^{N}(\hat{\theta}_j - \frac{1}{N}\sum_{h=1}^{N}\hat{\theta}_h)^2 \sum_{j=1}^{N}(\theta_j^{(l)} - \frac{1}{N}\sum_{h=1}^{N}\theta_h^{(l)})^2}$$

with $\theta_h^{(l)}$ denoting the value of θ_j on the *l*-th MCMC sample $\hat{\rho}^2_{(l)}$ denoting the posterior mean of $\sum_{l=1}^{5000}\theta_j^{(l)}/5000$ .. A 95% credible interval for ρ^2 was obtained by calculating the 125th smallest and 125th largest values of across the 5,000 MCMC samples. All hospitals regardless of sample size were included in the estimation of Bayesian model parameters. Reliability measures were initially calculated including all the hospitals and were subsequently calculated in subsets of hospitals with specified minimum number of performed procedures.

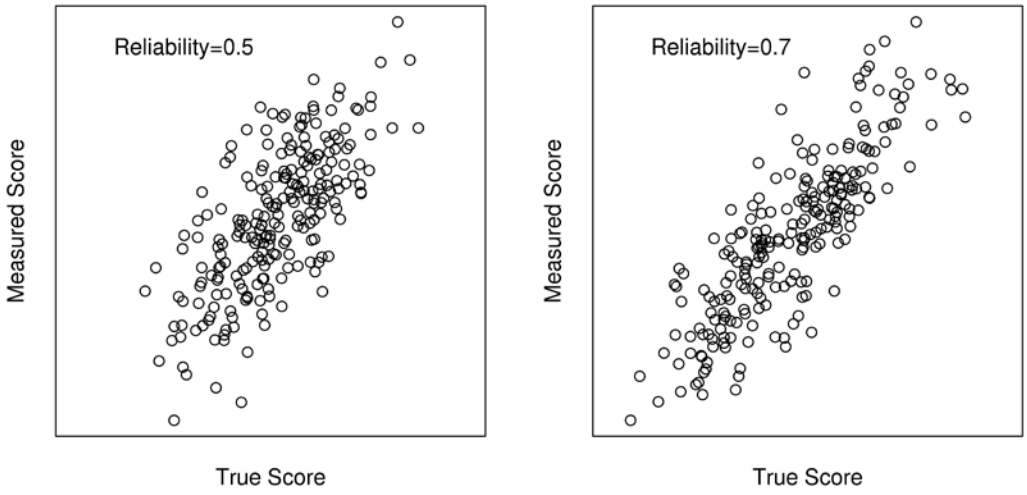
2a2.3. For each level of testing checked above, what were the statistical results from reliability testing? (e.*g., percent agreement and kappa for the critical data elements; distribution of reliability statistics from a signal-to-noise analysis*)
Prior to estimating reliability, the numerical value of SIR was estimated for each hospital under the model described by Fernandez et al. (2016). The reliability measure was calculated as the estimated squared correlation between the set of hospital-specific estimates of SIR and the corresponding unknown true values (estimated using Bayesian inference methods). A 95% Bayesian probability interval for this reliability measure was obtained. With all 231 hospitals included, the estimate of the reliability measure is 0.50 and the 95% Bayesian probability interval (0.42, 0.58), it is 0.53 (0.45, 0.61) for 216 hospitals performing at least 10 procedures, and it is 0.84 (0.76, 0.91) for 38 hospitals with 200 or more procedures performed.

Given the timeframe of the data used for reliability testing for this measure (01/01/2012 – 12/31/2014), the revised postoperative complication data element "unexpected return to the operating room" was included in the analysis.

2a2.4 What is your interpretation of the results in terms of demonstrating reliability? (i.*e.*, *what do the results mean and what are the norms for the test conducted?*)

In summary, when estimated with 3 years of data, the proposed lung cancer morbidity and mortality measure is reliable enough to be useful in the context of feedback reporting for internal quality improvement initiatives. Reliability increases when considering participants with increasing minimum number of cases. Starting with participants with at least 10 cases, there is a moderate reliability of 0.53, and reliability is 0.84 when only large-volume participants (at least 200 cases) are considered. The increase in reliability is the result of a more precise estimation of a participant's measure value; in other words with the same between-participants variability, the reliability increases when the participant measurement error decreases with more cases per participant.

To visualize this effect of a decreasing measurement error on reliability, while keeping the same between-participant variability, we created two figures illustrating the accuracy of the measured scores when the true reliability is 0.50 and 0.70. Because the true score for the composite measure is unknown, we used simulated data with formula Measured Score$_i$=True Score$_i$ + e_i where $i = 1,2,\ldots,231$ indicates the 231 participants and where True Score$_i$ and e_i both follow normal distributions. The standard deviations of the normal distributions were chosen such that the measure (score) has a reliability of 0.50 on the left figure and reliability of 0.70 on the right figure. Each figure has true score along the x-axis, and the estimated (measured) value of this true score along the y-axis. With a decreasing measurement error of the score (as is the case with increase in the number of cases per participant), the correlation between the true and measured values of the score increases, and thus also, equivalently, the reliability increases because reliability can be expressed as a square of this correlation (Pearson correlation). Although a high reliability of 0.70 shows a very close correlation between true and measured scores, a more moderate reliability of 0.50 still visualizes a strong association (correlation) between the true and measured values of the score.



2b1. VALIDITY TESTING

2b1.1. What level of validity testing was conducted? (*may be one or both levels*)

☐ **Critical data elements** (*data element validity must address ALL critical data elements*)

☐ **Performance measure score**

☒ **Empirical validity testing**

☐ **Systematic assessment of face validity of performance measure score as an indicator** of quality or resource use (*i.e., is an accurate reflection of performance on quality or resource use and can distinguish good from poor performance*) **NOTE**: Empirical validity testing is expected at time of maintenance review; if not possible, justification is required.

2b1.2. For each level of testing checked above, describe the method of validity testing and what it tests (*describe the steps—do not just name a method; what was tested, e.g., accuracy of data elements compared to authoritative source, relationship to another measure as expected; what statistical analysis was used)*
When data arrive at the data warehouse, they are checked carefully for logical inconsistencies, missing required fields, and parent/child variable relationship violations. Any inconsistencies or violations are communicated to participants in the detailed Data Quality Report that is generated automatically following each harvest file submission. Upon receipt of the Data Quality Report, participants are given an opportunity to correct the data, which substantially improves the quality and completeness of the data submitted for analysis. If the data inconsistencies are not changed by the participant prior to harvest close, the data warehouse performs consistency edits and/or parent/child edits on the data in order for them to be analyzable. Participants are informed of such edits to their data in the Data Quality Report.

Since 2010, the STS has contracted with Telligen (formerly IFMC) and, most recently, Cardiac Registry Support, LLC (CRS) to conduct audits of the STS General Thoracic Surgery Database on the Society´s behalf to evaluate the accuracy, consistency and comprehensiveness of data collection, which has validated the integrity of the data. Currently, auditors validate case inclusion and 15 lobectomy and 5 esophagectomy cancer cases are randomly chosen for review of 39 individual data elements. The auditors abstract each designated medical record to validate data elements previously submitted to the STS data warehouse. Agreement rates are calculated for each of the 39 elements as well as for an overall agreement rate. Five sites were randomly selected for the first audit, which took place in 2010. In 2016, 25 sites were audited.

2b1.3. What were the statistical results from validity testing? (*e.g., correlation; t-test*)
STS audited 10% of participants in the General Thoracic Surgery Database in 2016 using an independent auditing firm (CRS). The sites were randomly selected and audited for data completeness and accuracy. Auditors compared case logs at each facility and cases submitted to the STS GTSD to assess completeness of data submission. There was consistent agreement across all participants for data completeness. Data accuracy was assessed by reabstraction of 15 randomly chosen lobectomy cancer cases and 5 esophagectomy cancer cases, comparing 39 data elements in the medical chart with the data file submitted to the STS GTSD. The agreement rate was 96.78% for overall data accuracy in 2016, with a range in agreement from 94.3% to 99.0%.

For comparison, the overall agreement rates in 2010 and 2011 were 89.9% and 94.6%, respectively (across the 33 data elements reviewed at that time). The range in agreement was from 76.5% to 95.5% in 2010, and from 88.8% to 97.5% in 2011.

Aggregate agreement rates from the 2016 audit for each of the 39 variables (data elements) and for each of the variable categories are displayed in the table below. The STS does not have access to audit results at the level of individual surgical cases; we are therefore unable to provide the kappa statistic.

CATEGORY	FIELD_NAME	NUM	DEN	Agreement
PRE-OPERATIVE EVALUATION	OVERALL_ALL_FIELDS	6455	6738	95.80%
PRE-OPERATIVE EVALUATION	Admission Date	497	500	99.40%
PRE-OPERATIVE EVALUATION	Prior Cardiothoracic Surgery	488	500	97.60%
PRE-OPERATIVE EVALUATION	Pre-Op Chemo-Current Malignancy	489	500	97.80%

PRE-OPERATIVE EVALUATION	Pre-Op Thoracic Radiation Therapy	489	500	97.80%
PRE-OPERATIVE EVALUATION	Diabetes	413	423	97.64%
PRE-OPERATIVE EVALUATION	Diabetes Therapy	68	82	82.93%
PRE-OPERATIVE EVALUATION	Cigarette Smoking	489	500	97.80%
PRE-OPERATIVE EVALUATION	Pulmonary Function Tests Performed	419	423	99.05%
PRE-OPERATIVE EVALUATION	FEV1 Predicted	316	414	76.33%
PRE-OPERATIVE EVALUATION	Zubrod Score	491	500	98.20%
PRE-OPERATIVE EVALUATION	Lung Cancer	420	423	99.29%
PRE-OPERATIVE EVALUATION	Clinical Staging Method-Lung-EBUS	408	419	97.37%
PRE-OPERATIVE EVALUATION	Clinical Staging Method-Lung-PET or PET/CT	397	419	94.75%
PRE-OPERATIVE EVALUATION	Lung Cancer Tumor Size-T	377	419	89.98%
PRE-OPERATIVE EVALUATION	Lung Cancer Nodes-N	409	419	97.61%
PRE-OPERATIVE EVALUATION	Esophageal Cancer	77	77	100.00%
PRE-OPERATIVE EVALUATION	Clinical Staging Method-Esophageal-EUS	69	75	92.00%
PRE-OPERATIVE EVALUATION	Esophageal Cancer Tumor-T	68	72	94.44%
PRE-OPERATIVE EVALUATION	Clinical Diagnosis of Nodal Involvement	71	73	97.26%
DIAGNOSIS AND PROCEDURES	OVERALL_ALL FIELDS	4842	4978	97.27%
DIAGNOSIS AND PROCEDURES	Category of Disease-Primary	479	499	95.99%
DIAGNOSIS AND PROCEDURES	Date of Surgery	498	500	99.60%
DIAGNOSIS AND PROCEDURES	Procedure Start Time	493	500	98.60%
DIAGNOSIS AND PROCEDURES	Procedure End Time	482	500	96.40%
DIAGNOSIS AND PROCEDURES	ASA Classification	487	500	97.40%
DIAGNOSIS AND PROCEDURES	Procedure	500	500	100.00%
DIAGNOSIS AND PROCEDURES	Patient Disposition	491	500	98.20%
DIAGNOSIS AND PROCEDURES	Pathologic Staging-Lung Cancer-T	405	419	96.66%
DIAGNOSIS AND PROCEDURES	Pathologic Staging-Lung Cancer-N	411	419	98.09%
DIAGNOSIS AND PROCEDURES	Lung Cancer-Number of Nodes	385	419	91.89%
DIAGNOSIS AND PROCEDURES	Pathologic Staging-Esophageal Cancer-T	69	74	93.24%
DIAGNOSIS AND PROCEDURES	Pathologic Staging-Esophageal Cancer-N	73	74	98.65%
DIAGNOSIS AND PROCEDURES	Esophageal Cancer-Number of Nodes	69	74	93.24%
POST-OPERATIVE EVENTS	OVERALL_ALL FIELDS	1487	1500	99.13%
POST-OPERATIVE EVENTS	Unexpected Return to OR	493	500	98.60%
POST-OPERATIVE EVENTS	Pneumonia	494	500	98.80%
POST-OPERATIVE EVENTS	Initial Vent Support >48 Hours	500	500	100.00%
DISCHARGE	OVERALL_ALL FIELDS	1935	1993	97.09%
DISCHARGE	Discharge Date	499	500	99.80%

DISCHARGE	Discharge Status	490	500	98.00%
DISCHARGE	Readmission within 30 Days of Discharge	484	493	98.17%
DISCHARGE	Status 30 Days After Surgery	462	500	92.40%
	OVERALL_ALL FIELDS	**14719**	**15209**	**96.78%**

2b1.4. What is your interpretation of the results in terms of demonstrating validity? (i.*e., what do the results mean and what are the norms for the test conducted?*)
The most recent audits of the General Thoracic Surgery Database have demonstrated a high degree of data validity. Overall data accuracy rates have increased substantially since audits of the GTSD were first conducted in 2010; agreement ranges have also narrowed, indicating greater consistency in data accuracy among audited sites.

2b2. EXCLUSIONS ANALYSIS
NA ☐ no exclusions — *skip to section 2b4*

2b2.1. Describe the method of testing exclusions and what it tests (*describe the steps—do not just name a method; what was tested, e.g., whether exclusions affect overall performance scores; what statistical analysis was used*)

We excluded patients with missing age, sex, discharge mortality status, pathologic stage, and predicted forced expiratory volume in 1 second. In addition patients were excluded if they had an extrapleural pneumonectomy, completion pneumonectomy, carinal pneumonectomy, occult carcinoma or benign disease on final pathology, or an urgent, emergent, or palliative operation. We believe these are clinically appropriate exclusions and are necessary to make the measure a consistent performance measure for the comparison across participants. The exclusions are precisely defined and specified.

2b2.2. What were the statistical results from testing exclusions? (*include overall number and percentage of individuals excluded, frequency distribution of exclusions across measured entities, and impact on performance measure scores*)

There were 216 (0.7%) patients with extrapleural pneumonectomy, completion pneumonectomy, or carinal pneumonectomy; 156 (0.5%) patients with occult carcinoma or benign disease on final pathology; 3 (0.01%) with palliative operation (ASA VI); and 1510 (5.1%) non-elective status (urgent or emergent) operations, resulting in the overall exclusion of 6.3%. Impact of these exclusions on the performance measure is likely not meaningful due to a small number of cases excluded.

2b2.3. What is your interpretation of the results in terms of demonstrating that exclusions are needed to prevent unfair distortion of performance results? (*i.e., the value outweighs the burden of increased data collection and analysis. Note: **If patient preference is an exclusion**, the measure must be specified so that the effect on the performance score is transparent, e.g., scores with and without exclusion*)

For the measure to consistently quantify the surgical quality of lung resection for lung cancer per its definition, it is necessary to exclude patients if they had an extrapleural pneumonectomy, completion pneumonectomy, carinal pneumonectomy, occult carcinoma or benign disease on final pathology, or an urgent, emergent, or palliative operation.

2b3. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES
If not an intermediate or health outcome, or PRO-PM, or resource use measure, skip to section 2b5.

2b3.1. What method of controlling for differences in case mix is used?

☐ **No risk adjustment or stratification**

☒ **Statistical risk model with risk factors**

☐ **Stratification by risk categories**

☐ **Other,**

2b3.1.1 If using a statistical risk model, provide detailed risk model specifications, including the risk model method, risk factors, coefficients, equations, codes with descriptors, and definitions.

Bayesian hierarchical random-effects logistic regression modeling was used to estimate hospital-specific standardized incidence ratio (SIR) and a 95% Bayesian probability interval for SIR for each of 231 hospitals. Random-effects refers to the assumption that the provider-specific parameters of interest are assumed to arise from a specified distribution defined by parameters that are also estimated in the modelling process. This analytic method is the same method used in Fernandez, et al. (2016). Risk factors in the model were: age, sex, body mass index, hypertension, steroid therapy, congestive heart failure, coronary artery disease, peripheral vascular disease, reoperation, cerebrovascular disease, diabetes mellitus, forced expiratory volume in 1 second percent of predicted, induction therapy, renal dysfunction, cigarette smoking, Zubrod score, American Society of Anesthesiologists class, approach, pathologic stage, and procedure type.

Fernandez FG, Kosinski AS, Burfeind W, Park B, DeCamp MM, Seder C, Marshall B, Magee MJ, Wright CD, Kozower BD. The Society of Thoracic Surgeons Lung Cancer Resection Risk Model: Higher Quality Data and Superior Outcomes. Ann Thorac Surg. 2016 Aug;102(2):370-7.

2b3.2. If an outcome or resource use component measure is <u>not risk adjusted or stratified</u>, provide <u>rationale and analyses</u> to demonstrate that controlling for differences in patient characteristics (case mix) is not needed to achieve fair comparisons across measured entities.

n/a

2b3.3a. Describe the conceptual/clinical <u>and</u> statistical methods and criteria used to select patient factors (clinical factors or social risk factors) used in the statistical risk model or for stratification by risk (*e.g., potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of p<0.10; correlation of x or higher; patient factors should be present at the start of care*) **Also discuss any "ordering" of risk factor inclusion**; for example, are social risk factors added after all clinical factors?

Covariates in this model were selected a priori based on a combination of literature review and expert group consensus, and as described in Fernandez, et al. (2016). All covariates were retained in the model and were not added or removed based on a statistical variable selection algorithm.

No social risk factors were used in the statistical risk model or for stratification.

Fernandez FG, Kosinski AS, Burfeind W, Park B, DeCamp MM, Seder C, Marshall B, Magee MJ, Wright CD, Kozower BD. The Society of Thoracic Surgeons Lung Cancer Resection Risk Model: Higher Quality Data and Superior Outcomes. Ann Thorac Surg. 2016 Aug;102(2):370-7.

2b3.3b. How was the conceptual model of how social risk impacts this outcome developed? Please check all that apply:

☒ **Published literature**

☐ **Internal data analysis**

☒ **Other (please describe)**

Expert group consensus

2b3.4a. What were the statistical results of the analyses used to select risk factors?

Estimated odds ratios are summarized in the table below.

Variable	Composite Model (Mortality or Major Morbidity) OR (95% CI)	*p* Value
Age, 10-year increase	1.14 (1.08–1.90)	<0.001
Male	1.41 (1.29–1.53)	<0.001
Body mass index, kg/m^2		<0.001
≥18.5 to <25	1.00	
≥6.0 to <18.5	1.35 (1.09–1.66)	
≥25.0 to <30.0	0.83 (0.75–0.92)	
≥30.0 to <35.0	0.72 (0.63–0.82)	
≥35.0 to ≤99.9	0.83 (0.71–0.97)	
Hypertension	1.06 (0.96–1.16)	0.25
Steroids	1.33 (1.09–1.62)	0.005
Congestive heart failure	1.19 (0.97–1.46)	0.10
Coronary artery disease	1.14 (1.03–1.26)	0.011
Peripheral vascular disease	1.43 (1.26–1.63)	<0.001
Reoperation	1.32 (1.13–1.54)	<0.001
Cerebrovascular disease	1.11 (0.97–1.28)	0.14
Diabetes mellitus	1.01 (0.91–1.13)	0.84
% FEV_1, 10% decrease	1.12 (1.10–1.15)	<0.001
Induction therapy	1.20 (1.03–1.39)	0.022
Renal dysfunction	1.11 (0.84–1.46)	0.47
Cigarette smoking		<0.001
Never	1.00	
Past smoker	1.23 (1.05–1.44)	
Current smoker	1.64 (1.38–1.94)	
Zubrod score		<0.001
0	1.00	
1	1.16 (1.06–1.28)	
2–5	1.60 (1.32–1.95)	
ASA		<0.001
1 or 2	1.00	
3	1.27 (1.09–1.47)	
4 or 5	1.76 (1.45–2.13)	
Approach		<0.001
Minimally invasive	1.00	
Thoracotomy	1.51 (1.37–1.66)	
Pathologic stage		0.25
I	1.00	
II	1.05 (0.95–1.17)	
III	1.14 (1.00–1.30)	
IV	1.04 (0.75–1.42)	
Procedure		<0.001
Wedge	1.00	
Segmentectomy	1.24 (0.97–1.57)	
Lobectomy	1.93 (1.65–2.26)	
Sleeve	1.96 (1.39–2.77)	
Bilobectomy	2.91 (2.29–3.70)	
Pneumonectomy	2.83 (2.24–3.58)	

Fernandez FG, Kosinski AS, Burfeind W, Park B, DeCamp MM, Seder C, Marshall B, Magee MJ, Wright CD, Kozower BD. The Society of Thoracic Surgeons Lung Cancer Resection Risk Model: Higher Quality Data and Superior Outcomes. Ann Thorac Surg. 2016 Aug;102(2):370-7.

2b3.4b. Describe the analyses and interpretation resulting in the decision to select social risk factors *(e.g. prevalence of the factor across measured entities, empirical association with the outcome, contribution of unique variation in the outcome, assessment of between-unit effects and within-unit effects.)* **Also describe the impact of adjusting for social risk (or not) on providers at high or low extremes of risk.**

All covariates were retained in the model and were not added or removed based on a statistical variable selection algorithm.

As noted in 1.8 above, patient social risk data are not collected in the General Thoracic Surgery Database.

2b3.5. Describe the method of testing/analysis used to develop and validate the adequacy of the statistical model or stratification approach (*describe the steps—do not just name a method; what statistical analysis was used*)

Continuous variables were evaluated with respect to linearity of effect and no departure from linearity was noted. The calibration of the model was assessed with the Hosmer-Lemeshow goodness-of-fit statistic. The discrimination of the model was assessed with the C-statistic.

Provide the statistical results from testing the approach to controlling for differences in patient characteristics (case mix) below.

If stratified, skip to 2b3.9

2b3.6. Statistical Risk Model Discrimination Statistics (*e.g., c-statistic, R-squared*)**:**

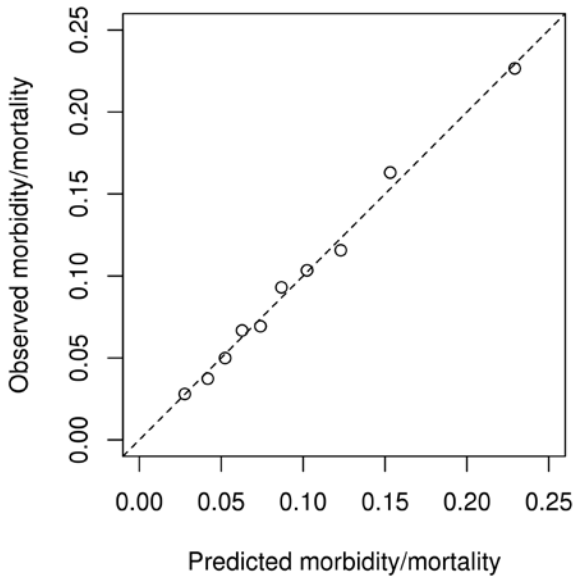
The C-statistics is 0.68.

2b3.7. Statistical Risk Model Calibration Statistics (*e.g., Hosmer-Lemeshow statistic*):

The Hosmer-Lemeshow goodness-of-fit p-value=0.40 demonstrates that the model estimates fit the data at an acceptable level.

2b3.8. Statistical Risk Model Calibration – Risk decile plots or calibration curves:

Risk decile plot below shows good alignment of predicted and observed probabilities of outcome (operative mortality or major morbidity) within deciles of predicted values.



2b3.9. Results of Risk Stratification Analysis:

n/a

2b3.10. What is your interpretation of the results in terms of demonstrating adequacy of controlling for differences in patient characteristics (case mix)? (i.*e., what do the results mean and what are the norms for the test conducted*)

The results demonstrated that the STS lung resection for lung cancer risk model is well calibrated and has good discrimination power. It is suitable for controlling for differences in case-mix between centers.

2b3.11. Optional Additional Testing for Risk Adjustment (*not required, but would provide additional support of adequacy of risk model, e.g., testing of risk model in another data set; sensitivity analysis for missing data; other methods that were assessed*)
n/a

2b4. IDENTIFICATION OF STATISTICALLY SIGNIFICANT & MEANINGFUL DIFFERENCES IN PERFORMANCE

2b4.1. Describe the method for determining if statistically significant and clinically/practically meaningful differences in performance measure scores among the measured entities can be identified (*describe the steps—do not just name a method; what statistical analysis was used? Do not just repeat the information provided related to performance gap in 1b)*

Bayesian hierarchical modeling was used to estimate hospital-specific standardized incidence ratio (SIR) and a 95% Bayesian probability interval for SIR for each of 231 hospitals. The degree of uncertainty surrounding an STS participant's SIR is indicated by calculating 95% Bayesian credible intervals (CrI's) which are similar to conventional confidence intervals. An STS participant's performance is considered average if the Bayesian credible interval (CrI) surrounding their SIR score overlaps 1. If the Bayesian CrI falls entirely below 1, the participant has lower-than-expected performance. If the Bayesian CrI falls entirely above 1, the participant has higher-than-expected performance.

2b4.2. What were the statistical results from testing the ability to identify statistically significant and/or clinically/practically meaningful differences in performance measure scores across measured entities? (e.g., *number and percentage of entities with scores that were statistically significantly different from mean or some benchmark, different from expected; how was meaningful difference defined*)

Figure 1 under the Results section of the attachment (Fernandez et al, 2016) displays estimated SIR and corresponding 95% Bayesian probability interval for each of 231 hospitals. Hospitals are ordered according to the increasing SIR estimate. There are meaningful differences between the best performing (3.5%; 8 of 231 sites) and the worst performing hospitals (6.9%; 16 of 231 sites). This indicates that this model provides meaningful discrimination between best and worst performers.

2b4.3. What is your interpretation of the results in terms of demonstrating the ability to identify statistically significant and/or clinically/practically meaningful differences in performance across measured entities? (i.*e., what do the results mean in terms of statistical and meaningful differences?*)

The identified differences in performance between centers are both statistically significant and clinically meaningful. The surgeon panel and users are satisfied with the distribution of participants across performance categories.

2b5. COMPARABILITY OF PERFORMANCE SCORES WHEN MORE THAN ONE SET OF SPECIFICATIONS

If only one set of specifications, this section can be skipped.

Note: *This item is directed to measures that are risk-adjusted (with or without social risk factors)* ***OR*** *to measures with more than one set of specifications/instructions (e.g., one set of specifications for how to identify and compute the measure from medical record abstraction and a different set of specifications for claims or eMeasures). It does not apply to measures that use more than one source of data in one set of specifications/instructions (e.g., claims data to identify the denominator and medical record abstraction for the numerator).* ***Comparability is not required when comparing performance scores with and without social risk factors in the risk adjustment model. However, if comparability is not demonstrated for measures with more than one set of specifications/instructions, the different specifications (e.g., for medical records vs. claims) should be submitted as separate measures.***

2b5.1. Describe the method of testing conducted to compare performance scores for the same entities across the different data sources/specifications (*describe the steps—do not just name a method; what statistical analysis was used*)
n/a

2b5.2. What were the statistical results from testing comparability of performance scores for the same entities when using different data sources/specifications? (*e.g., correlation, rank order*)
n/a

2b5.3. What is your interpretation of the results in terms of the differences in performance measure scores for the same entities across the different data sources/specifications? (i.*e., what do the results mean and what are the norms for the test conducted*)
n/a

2b6. MISSING DATA ANALYSIS AND MINIMIZING BIAS

2b6.1. Describe the method of testing conducted to identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias (*describe the steps—do not just name a method; what statistical analysis was used*)

The quality of data in STS General Thoracic Surgery Database has been improving. We managed the remaining missing data with imputation. Missing body mass index (BMI) values (1%) were imputed utilizing sex specific median of the observed BMI values. For binary risk factors, missing values were considered as indicating absence of the risk factor.

2b6.2. What is the overall frequency of missing data, the distribution of missing data across providers, and the results from testing related to missing data? (*e.g., results of sensitivity analysis of the effect of various rules for missing data/nonresponse; if no empirical sensitivity analysis, identify the approaches for handling missing data that were considered and pros and cons of each*)

Patients with missing age, sex, discharge mortality status, pathologic stage, and predicted forced expiratory volume in 1 second were excluded. All the variables in the population utilized for this measure had less than 1% of missing values.

2b6.3. What is your interpretation of the results in terms of demonstrating that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias**?** (i.*e.*, *what do the results mean in terms of supporting the selected approach for missing data and what are the norms for the test conducted; if no empirical analysis, provide rationale for the selected approach for missing data*)

The rates of missing data were low. We therefore concluded that systematic missing data did not lead to bias in our measure.

3. Feasibility

Extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

3a. Byproduct of Care Processes

For clinical measures, the required data elements are routinely generated and used during care delivery (e.g., blood pressure, lab test, diagnosis, medication order).

3a.1. **Data Elements Generated as Byproduct of Care Processes.**
generated by and used by healthcare personnel during the provision of care, e.g., blood pressure, lab value, medical condition, Coded by someone other than person obtaining original information (e.g., DRG, ICD-9 codes on claims), Abstracted from a record by someone other than person obtaining original information (e.g., chart abstraction for quality measure or registry)
If other:

3b. Electronic Sources

The required data elements are available in electronic health records or other electronic sources. If the required data are not in electronic health records or existing electronic sources, a credible, near-term path to electronic collection is specified.

3b.1. **To what extent are the specified data elements available electronically in defined fields** (*i.e., data elements that are needed to compute the performance measure score are in defined, computer-readable fields*) Update this field for **maintenance of endorsement**.
ALL data elements are in defined fields in a combination of electronic sources

3b.2. **If ALL the data elements needed to compute the performance measure score are not from electronic sources, specify a credible, near-term path to electronic capture, OR provide a rationale for using other than electronic sources.** For **maintenance of endorsement**, if this measure is not an eMeasure (eCQM), please describe any efforts to develop an eMeasure (eCQM).
n/a

3b.3. **If this is an eMeasure, provide a summary of the feasibility assessment in an attached file or make available at a measure-specific URL. Please also complete and attach the NQF Feasibility Score Card.**
Attachment:

3c. Data Collection Strategy

Demonstration that the data collection strategy (e.g., source, timing, frequency, sampling, patient confidentiality, costs associated with fees/licensing of proprietary measures) can be implemented (e.g., already in operational use, or testing demonstrates that it is ready to put into operational use). For eMeasures, a feasibility assessment addresses the data elements and measure logic and demonstrates the eMeasure can be implemented or feasibility concerns can be adequately addressed.

3c.1. **Required for maintenance of endorsement. Describe difficulties (as a result of testing and/or operational use of the measure) regarding data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, time and cost of data collection, other feasibility/implementation issues.**
IF instrument-based, consider implications for both individuals providing data (patients, service recipients, respondents) and those whose performance is being measured.
Missing data are sought by the DCRI from participants when the data are initially sent to DCRI for analysis.

Data are collected continuously by the participating sites and harvested by the DCRI twice yearly. Reports are then sent back to the sites about 3 months after a harvest.

No individual patient identifiers are collected by the DCRI.

Data Collection:
Participants of the STS General Thoracic Surgery Database generally have data managers on staff to collect these data. Costs to develop the measure included volunteer thoracic surgeons' time, STS staff time, and DCRI statistician and project management time.

Other fees:
STS General Thoracic Surgery Database participant surgeons pay an annual participant fee of $550 or $700, depending on whether the participant is an STS member or not. STS membership thus provides surgeons with a 21% discount on the non-member database participation fee.

3c.2. **Describe any fees, licensing, or other requirements to use any aspect of the measure as specified** *(e.g., value/code set, risk model, programming code, algorithm)*.
See 3c.1

4. Usability and Use

Extent to which potential audiences (e.g., consumers, purchasers, providers, policy makers) are using or could use performance results for both accountability and performance improvement to achieve the goal of high-quality, efficient healthcare for individuals or populations.

4a. Accountability and Transparency

Performance results are used in at least one accountability application within three years after initial endorsement and are publicly reported within six years after initial endorsement (or the data on performance results are available). If not in use at the time of initial endorsement, then a credible plan for implementation within the specified timeframes is provided.

4.1. **Current and Planned Use**
NQF-endorsed measures are expected to be used in at least one accountability application within 3 years and publicly reported within 6 years of initial endorsement in addition to performance improvement.

Specific Plan for Use	Current Use (for current use provide URL)
Public Reporting	Quality Improvement (external benchmarking to organizations) STS General Thoracic Surgery Database http://publicreporting.sts.org/gtsd Quality Improvement (Internal to the specific organization) STS General Thoracic Surgery Database http://publicreporting.sts.org/gtsd

4a1.1 **For each CURRENT use, checked above (update for maintenance of endorsement), provide:**

- Name of program and sponsor
- Purpose
- Geographic area and number and percentage of accountable entities and patients included
- Level of measurement and setting

See 4a1.2

4a1.2. **If not currently publicly reported OR used in at least one other accountability application (e.g., payment program, certification, licensing) what are the reasons?** (*e.g., Do policies or actions of the developer/steward or accountable entities restrict access to performance results or impede implementation?*)

STS is actively promoting public reporting of the STS adult cardiac, congenital heart, and general thoracic surgery performance measures. This is consistent with the explicitly stated STS philosophy that "As a national leader in health care transparency and accountability, The Society of Thoracic Surgeons believes that the public has a right to know the quality of surgical outcomes." (http://www.sts.org/registries-research-center/sts-public-reporting) In our efforts to operationalize public reporting, the STS Public Reporting Task Force has and will continue to develop public report cards that are consumer centric. Public reporting remains a top priority for the Society, and STS is striving for even stronger involvement among Database participants.

Currently, more than 650 Adult Cardiac Surgery Database (ACSD) participants voluntarily consent to be a part of the STS Public Reporting and more than 550 ACSD participants have consented to report publicly via the Consumer Reports public reporting initiative. Additionally, more than 100 Congenital Heart Surgery Database (CHSD) participants are currently enrolled in STS Public Reporting.

As of July 2017, General Thoracic Surgery Database (GTSD) participants were included in the Public Reporting initiative and more than 250 participants currently consent to report outcomes publicly on the STS website. This includes discharge mortality rate and median postoperative length of stay for lobectomy procedures for lung cancer, including scores and star ratings for the Lobectomy for Lung Cancer Composite Measure in addition to its domains of 1) absence of mortality, and 2) absence of major complication. Participant outcomes are published alongside GTSD overall outcomes and National Inpatient Sample (NIS) outcomes.

-ACSD public reporting online may be found here: http://publicreporting.sts.org/acsd
-CHSD public reporting online may be found here: http://publicreporting.sts.org/chsd
-GHSD public reporting online may be found here: http://publicreporting.sts.org/gtsd

4a1.3. **If not currently publicly reported OR used in at least one other accountability application, provide a credible plan for implementation within the expected timeframes -- any accountability application within 3 years and publicly reported within 6 years of initial endorsement.** (*Credible plan includes the specific program, purpose, intended audience, and timeline for implementing the measure within the specified timeframes. A plan for accountability applications addresses mechanisms for data aggregation and reporting.*)
See 4a1.2

4a2.1.1. **Describe how performance results, data, and assistance with interpretation have been provided to those being measured or other users during development or implementation.**
How many and which types of measured entities and/or others were included? If only a sample of measured entities were included, describe the full population and how the sample was selected.
STS's combined mortality and morbidity model for pulmonary resection for lung cancer is important and appropriate for public reporting for the following reasons:
1.) lung cancer resection is the most common category of surgical procedures that a thoracic surgeon performs;
2.) these procedures are therefore useful and appropriate to use as a benchmark for performance by general thoracic surgery programs. By providing surgeons and teams with risk-adjusted results, they can identify how they are performing compared with other programs in the STS General Thoracic Database, which generally includes the top thoracic programs in the nation. This will assist them in focusing performance improvement efforts. Also, when publicly reported, the outcomes for these common procedures provide patients and their families with comparative performance information to aid in selection of a provider;
3.) major morbidity is relatively common after lung resection; however, although mortality is rare, it should be captured as well in an outcome measure, thereby identifying ALL adverse events after lung resection;
4.) this measure is reported in an easy to understand format which summarizes the results of all participants who were included in the analysis. The participant's score is illustrated graphically in relation to the 25th, 50th and 75th percentiles of the distribution across participants, and is accompanied by the 95% Bayesian credible interval. Surgeons easily grasp this result and the visual display powerfully shows them just where they perform compared to their peers on a bi-annual basis. In addition, these risk-adjusted results allow surgeons to compare their patients´ outcomes with national benchmarks and to initiate QI efforts as needed.

4a2.1.2. **Describe the process(es) involved, including when/how often results were provided, what data were provided, what educational/explanatory efforts were made, etc.**
See 4a2.1.1

4a2.2.1. **Summarize the feedback on measure performance and implementation from the measured entities and others described in 4d.1.**
Describe how feedback was obtained.

The general thoracic surgeons from across the U.S. who comprise the STS General Thoracic Surgery Task Force meet periodically to discuss the participant reports and to consider potential enhancements to the GTSD. Additions/clarifications to the data collection form and to the content/format of the participant reports are discussed and implemented as appropriate.

Most recently, STS surgeon members have expressed interest in real-time, online data updates, which has led to the development of dashboard-type reporting on STS.org. The general thoracic dashboard is scheduled for launch in 2018.

Also, general thoracic public reporting was initiated in the summer of 2017 (http://publicreporting.sts.org/gtsd), making star ratings for consenting participant groups available to participants as well as the public.

4a2.2.2. **Summarize the feedback obtained from those being measured.**
See 4a2.2.1

4a2.2.3. **Summarize the feedback obtained from other users**
Given the very recent launch of general thoracic public reporting, the STS has not yet received sufficient feedback from non-participants to be able to assess the impact of the public reporting initiative.

4a2.3. **Describe how the feedback described in 4a2.2.1 has been considered when developing or revising the measure specifications or implementation, including whether the measure was modified and why or why not.**
See Specifications section, S.3.2, regarding modification in postoperative complications included in numerator since most recent NQF review of this measure.

Improvement
Progress toward achieving the goal of high-quality, efficient healthcare for individuals or populations is demonstrated. If not in use for performance improvement at the time of initial endorsement, then a credible rationale describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.

4b1. **Refer to data provided in 1b but do not repeat here. Discuss any progress on improvement (trends in performance results, number and percentage of people receiving high-quality healthcare; Geographic area and number and percentage of accountable entities and patients included.)**
If no improvement was demonstrated, what are the reasons? If not in use for performance improvement at the time of initial endorsement, provide a credible rationale that describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.
Operative mortality in the STS General Thoracic Surgery Database has decreased from 2.2% in the years 2002 to 2008 to 1.4% from 2012 to 2014. These data represent the highest quality lung cancer surgery in the United States. It is important to recognize that a large proportion of the general thoracic surgery in the US is not performed by general thoracic surgeons certified by the American Board of Thoracic Surgery. Results by STS General Thoracic Database participants, who are almost all ABTS certified, are generally superior to those of surgeons performing these procedures who do not participate in the GTSD, and who are often not ABTS certified.

Kozower and colleagues (Ann Thorac Surg 2010) have previously demonstrated that compared with the Nationwide Inpatient Sample database, from 2002 to 2008, patients in the GTSD had lower unadjusted discharge mortality rates, median length of stay, and pulmonary complication rates for lobectomy.

The major morbidity rate has increased from 8.6% to 9.1% during the same time. A potential explanation for this observation is more complete coding of complications by data abstractors as the result of education efforts from STS, as well as inclusion of unexpected return to the operating room for any reason instead of only for bleeding.

Fernandez FG, Kosinski AS, Burfeind W, Park B, DeCamp MM, Seder C, Marshall B, Magee MJ, Wright CD, Kozower BD. The Society of Thoracic Surgeons Lung Cancer Resection Risk Model: Higher Quality Data and Superior Outcomes. Ann Thorac Surg. 2016 Aug;102(2):370-7.

Kozower BD, Sheng S, O'Brien SM, et al. STS database risk models: predictors of mortality and major morbidity for lung cancer resection. Ann Thorac Surg 2010;90:875–83.

4b2. Unintended Consequences

The benefits of the performance measure in facilitating progress toward achieving high-quality, efficient healthcare for individuals or populations outweigh evidence of unintended negative consequences to individuals or populations (if such evidence exists).

4b2.1. **Please explain any unexpected findings (positive or negative) during implementation of this measure including unintended impacts on patients.**
We are not aware of any unexpected findings associated with implementation of this measure.

4b2.2. **Please explain any unexpected benefits from implementation of this measure.**
n/a

5. Comparison to Related or Competing Measures

If a measure meets the above criteria and there are endorsed or new related measures (either the same measure focus or the same target population) or competing measures (both the same measure focus and the same target population), the measures are compared to address harmonization and/or selection of the best measure.

5. Relation to Other NQF-endorsed Measures
Are there related measures (conceptually, either same measure focus or target population) or competing measures (conceptually both the same measure focus and same target population)? If yes, list the NQF # and title of all related and/or competing measures.
Yes

5.1a. **List of related or competing measures (selected from NQF-endorsed measures)**

5.1b. **If related or competing measures are not NQF endorsed please indicate measure title and steward.**
This measure is related conceptually to the STS Lobectomy for Lung Cancer Composite Score measure, which we are submitting for initial NQF review in the fall 2017 Surgery endorsement cycle. The numerators for both measures include the same list of postoperative complications, but the outcomes for the Lobectomy Composite measure are grouped into two domains (operative mortality and major complications) and the measure is structured to provide general thoracic surgeons with a "star rating." Please also see 5a.2 below.

5a. Harmonization of Related Measures
The measure specifications are harmonized with related measures;
OR
The differences in specifications are justified

5a.1. **If this measure conceptually addresses EITHER the same measure focus OR the same target population as NQF-endorsed measure(s):**
Are the measure specifications harmonized to the extent possible?
Yes

5a.2. **If the measure specifications are not completely harmonized, identify the differences, rationale, and impact on interpretability and data collection burden.**
Measure #1790 includes a broader range of lung resection procedures than the Lobectomy Composite, and therefore includes a larger number of cases and potentially provides performance data to more general thoracic surgeons. Of the two measures, only the Lobectomy Composite is currently publicly reported.

5b. Competing Measures
The measure is superior to competing measures (e.g., is a more valid or efficient way to measure);
OR
Multiple measures are justified.

5b.1. **If this measure conceptually addresses both the same measure focus and the same target population as NQF-endorsed measure(s):**

Describe why this measure is superior to competing measures (e.g., a more valid or efficient way to measure quality); OR provide a rationale for the additive value of endorsing an additional measure. (Provide analyses when possible.)
n/a

Appendix

A.1 Supplemental materials may be provided in an appendix. All supplemental materials (such as data collection instrument or methodology reports) should be organized in one file with a table of contents or bookmarks. If material pertains to a specific submission form number, that should be indicated. Requested information should be provided in the submission form and required attachments. There is no guarantee that supplemental materials will be reviewed.
Attachment **Attachment:** FernandezKosinskiKozower_lung_cancer_risk_model_2016.pdf

Contact Information

Co.1 **Measure Steward (Intellectual Property Owner):** The Society of Thoracic Surgeons
Co.2 **Point of Contact:** Mark, Antman, mantman@sts.org, 312-202-5856-
Co.3 **Measure Developer if different from Measure Steward:** The Society of Thoracic Surgeons
Co.4 **Point of Contact:** Mark, Antman, mantman@sts.org, 312-202-5856-

Additional Information

Ad.1 **Workgroup/Expert Panel involved in measure development**
Provide a list of sponsoring organizations and workgroup/panel members' names and organizations. Describe the members' role in measure development.
Members of the STS Task Force on Quality Initiatives provide surgical expertise as needed. The STS Workforce on National Databases meets at the STS Annual Meeting and reviews the measures on a yearly basis. Changes or updates to the measure will be at the recommendation of the Workforce.

Measure Developer/Steward Updates and Ongoing Maintenance
Ad.2 **Year the measure was first released:** 2010
Ad.3 **Month and Year of most recent revision:** 02, 2016
Ad.4 **What is your frequency for review/update of this measure?** annually
Ad.5 **When is the next scheduled review/update for this measure?** 01, 2018

Ad.6 **Copyright statement:**
Ad.7 **Disclaimers:**

Ad.8 **Additional Information/Comments:**