NATIONAL QUALITY FORUM
Driving measurable health improvements together

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 #: 0697

Corresponding Measures:

Measure Title: Risk Adjusted Case Mix Adjusted Elderly Surgery Outcomes Measure

Measure Steward: American College of Surgeons

sp.02. Brief Description of Measure: This is a hospital based, risk adjusted, case mix adjusted elderly surgery aggregate clinical outcomes measure of adults 65 years of age and older.

1b.01. Developer Rationale:

2017 Maintenance

Reduced mortality and major morbidity rates for elderly following surgeries.

---

sp.12. Numerator Statement:

2022 Maintenance

This measure examines the occurrence of a mortality/morbidity composite defined later in the submission. NSQIP routinely adjusts definitions of outcomes for reasons that could include: Enhancement of clinical meaningfulness, simplicity, concordance with definitions constructed by other entities, and so forth. For the most part, these changes are minor and have limited impact on which cases are defined as having experienced an event or not. Furthermore, as revised definitions apply to all patients at participating hospitals, these changes do not impact the fairness of risk adjustment.

2022 definitions are described in section sp13 in reference to 2016 definitions.

2017 Maintenance

The outcome of interest is hospital-specific risk-adjusted mortality, a return to the operating room, or any of the following morbidities as defined by American College of Surgeons National Surgical Quality Improvement Program (ACS NSQIP): Cardiac Arrest requiring CPR, Myocardial Infarction, Sepsis, Septic Shock, Deep Incisional Surgical Site Infection (SSI), Organ/Space SSI, Wound Disruption, Unplanned Reintubation without prior ventilator dependence, Pneumonia without pre-operative pneumonia, progressive Renal Insufficiency or Acute Renal Failure without pre-operative renal failure or dialysis, or urinary tract infection (UTI) within 30 days of any ACS NSQIP listed (CPT) surgical procedure. The original endorsed measure included venous thromboembolism (VTE) as eligible morbidity events, including deep venous thrombosis requiring therapy and pulmonary embolism.

sp.14. Denominator Statement:

2017 Maintenance

Patients undergoing any ACS NSQIP listed (CPT) surgical procedure who are 65 years of age or older. (See appendix of roughly 2900 ACS NSQIP eligible CPT codes)

sp.16. Denominator Exclusions:

2017 Maintenance

Cases must first have ACS NSQIP eligible CPT codes on the submitted list.

2017 Maintenance

Cases must first have ACS NSQIP eligible CPT codes on the submitted list.

Measure Type: Outcome

sp.28. Data Source:

Other

Registry Data

Electronic Health Records

Management Data

Electronic Health Data

Paper Medical Records

sp.07. Level of Analysis:

Facility

IF Endorsement Maintenance – Original Endorsement Date: 2011-01-17 12:00 AM

Most Recent Endorsement Date: 1/25/2017 1:17:20 PM

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

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

sp.03. IF PAIRED/GROUPED, what is the reason this measure must be reported with other measures to appropriately interpret results?:

1. Importance to Measure and Report

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

---

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

[Response Begins]

No

[Response Ends]

Please separate added or updated information from the most recent measure evaluation within each question response in the Importance to Measure and Report: Evidence section. For example:

Current Submission:

Updated evidence information here.

Previous (Year) Submission:

Evidence from the previous submission here.

1a.01. Provide a logic model.

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

[Response Begins]

2017 Maintenance

Identification of poor performance on the measure identifies the potential to improve and be comparable to better performing hospitals. NSQIP provides documentation of structures and processes that, when implemented, have been shown to improve outcomes.

[Response Ends]

1a.02. Provide evidence that the target population values the measured outcome, process, or structure and finds it meaningful.

Describe how and from whom input was obtained.

[Response Begins]

[Response Ends]

1a.03. Provide empirical data demonstrating the relationship between the outcome (or PRO) and at least one healthcare structure, process, intervention, or service.

[Response Begins]

2022 maintenance

The Measure Elderly model is one outcome tracked by NSQIP that hospitals use to direct QI projects and improve surgical outcomes. If one examines Measure Elderly DSM, there is a trend towards fewer adverse events from 2016 through 2021 (raw DSM rates = 10.82, 9.57, 9.25, 9.04, 8,87, 9.81, respectively). The increase in 2021 may reflect the fact that there were fewer, but a greater percentage of emergent, higher risk procedures conducted during the COVID pandemic and with some adverse events being the result of pre- or post-operative COVD.

2017 Maintenance

The rates of the serious events described in this measure are highly variable by institution. ACS NSQIP uses clinical, audited, third - party collection, and risk adjusted data. Over time, performance has improved for hospitals participating in NSQIP. The majority of hospitals experience declines in mortality and morbidity, with annual reductions of approximately 0.8% and 3.1%, respectively. (Cohen, Liu et al. 2016) For 2014, there were 460 hospitals contributing 206,064 surgical cases on adults age 65 and over. The O/E ratios for mortality and serious morbidity in the elderly (age equal or greater than 65 years) range from 0.59 to 1.69 for participating hospitals. The interquartile range for O/E ratios is 0.23 and the 10th percentile and 90th percentile O/E ratios were 0.79 and 1.22, respectively. These statistics demonstrate the significance of the performance gap in mortality and serious morbidity outcomes in the elderly across hospitals.

Cohen, M. E., Y. Liu, C. Y. Ko and B. L. Hall. Improved surgical outcomes for ACS NSQIP hospitals over time – evaluation of hospitalcohorts with up to 8 years of participation. Ann Surg. 2016; 263:267-273

[Response Ends]

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

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

[Response Begins]

2017 Maintenance

Reduced mortality and major morbidity rates for elderly following surgeries.

[Response Ends]

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

Include mean, std dev, min, max, interquartile range, and scores by decile. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities include. This information also will be used to address the sub-criterion on improvement (4b) under Usability and Use.

[Response Begins]

2017 Maintenance

The rates of the serious events described in this measure are highly variable by institution. ACS NSQIP uses clinical, audited, third -party collection, and risk adjusted data. Over time, performance has improved for hospitals participating in NSQIP. The majority of hospitals experience declines in mortality and morbidity, with annual

reductions of approximately 0.8% and 3.1%, respectively. (Cohen, Liu et al. 2016) For 2014, there were 460 hospitals contributing 206,064 surgical cases on adults age 65 and over. The O/E ratios for mortality and serious morbidity in the elderly (age equal or greater than 65 years) range from 0.59 to 1.69 for participating hospitals. The interquartile range for O/E ratios is 0.23 and the 10th percentile and 90th percentile O/E ratios were 0.79 and 1.22, respectively. These statistics demonstrate the significance of the performance gap in mortality and serious morbidity outcomes in the elderly across hospitals.

Cohen, M. E., Y. Liu, C. Y. Ko and B. L. Hall. Improved surgical outcomes for ACS NSQIP hospitals over time – evaluation of hospital cohorts with up to 8 years of participation. Ann Surg. 2016; 263:267-273

[Response Ends]

1b.03. If no or limited performance data on the measure as specified is reported above, then provide a summary of data from the literature that indicates opportunity for improvement or overall less than optimal performance on the specific focus of measurement. Include citations.

[Response Begins]

2017 Maintenance

The data cited above is unpublished, obtained from an internal analysis of ACS NSQIP data. However, these gaps have been repeatedly demonstrated since the inception of the program.

[Response Ends]

1b.04. Provide disparities data from the measure as specified (current and over time) by population group, e.g., by race/ethnicity, gender, age, insurance status, socioeconomic status, and/or disability.

Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included. Include mean, std dev, min, max, interquartile range, and scores by decile. For measures that show high levels of performance, i.e., "topped out", disparities data may demonstrate an opportunity for improvement/gap in care for certain sub-populations. This information also will be used to address the sub-criterion on improvement (4b) under Usability and Use.

[Response Begins]

2017 Maintenance

For older adults (those aged 65 years and older), there are dramatic variations in the quality and delivery of surgical care (Dunlop et al. 2008; Laycock et al. 2000; Wanebo et al. 1997) as well as inclusion in clinical trials (Murthy et al. 2004; Bugeja et al. 1997). Birkmeyer et al. demonstrated that Medicare beneficiaries undergoing one of six major surgical procedures who belonged to a lower socioeconomic class had higher rates of adjusted mortality than those from a higher class, attributing the variation in outcomes to hospital-level differences in care.(Birkmeyer et al. 2008) Furthermore, in the Nationwide Inpatient Sample, operative mortality among patients aged 65 years and older who underwent pancreatic resection and esophagectomy was 10% less at high-volume centers compared to low-volume centers.(Finlayson, Birkmeyer 2001) Hardiman et al. demonstrated that among 10,433 patients diagnosed with primary colon tumors, individuals who were 80 years or older were less likely to have colectomy for advanced or metastatic disease, have fewer lymph nodes removed, and receive chemotherapy for every stage than those younger than 80 years old.(Hardiman et al. 2009) Skinner et al. found that the rate of surgical treatment of osteoarthritis of the knee in Medicare beneficiaries varies substantially by region of the country, sex, and race or ethnicity.(Skinner et al. 2003) Jha et al. confirmed the persistence of significant racial disparities in the performance of coronary artery bypass grafts, carotid endarterectomy, and total hip replacement among Medicare beneficiaries despite federal initiatives to reduce this variation.(Jha et al. 2005)

[Response Ends]

1b.05. If no or limited data on disparities from the measure as specified is reported above, then provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement. Include citations. Not necessary if performance data provided in above.

[Response Begins]

Birkmeyer NJ, Gu N, Baser O, Morris AM, Birkmeyer JD. Socioeconomic status and surgical mortality in the elderly. Med Care. Sep 2008;46(9):893-899.
Bugeja G, Kumar A, Banerjee AK. Exclusion of elderly people from clinical research: a descriptive study of published reports. BMJ. Oct 25 1997;315(7115):1059.
Dunlop DD, Manheim LM, Song J, et al. Age and racial/ethnic disparities in arthritis-related hip and knee surgeries. Med Care. Feb 2008;46(2):200-208.
Finlayson EV, Birkmeyer JD. Operative mortality with elective surgery in older adults. Eff Clin Pract. Jul-Aug 2001;4(4):172-177.
Hardiman KM, Cone M, Sheppard BC, Herzig DO. Disparities in the treatment of colon cancer in octogenarians. Am J Surg. May 2009;197(5):624-628.
Jha AK, Fisher ES, Li Z, Orav EJ, Epstein AM. Racial trends in the use of major procedures among the elderly. N Engl J Med. Aug 18 2005;353(7):683-691.
Laycock WS, Siewers AE, Birkmeyer CM, Wennberg DE, Birkmeyer JD. Variation in the use of laparoscopic cholecystectomy for elderly patients with acute cholecystitis. Arch Surg. Apr 2000;135(4):457-462.
Murthy VH, Krumholz HM, Gross CP. Participation in cancer clinical trials: race-, sex-, and age-based disparities. JAMA. Jun 9 2004;291(22):2720-2726.
Skinner J, Weinstein JN, Sporer SM, Wennberg JE. Racial, ethnic, and geographic disparities in rates of knee arthroplasty among Medicare patients. N Engl J Med. Oct 2 2003;349(14):1350-1359.
Wanebo HJ, Cole B, Chung M, et al. Is surgical management compromised in elderly patients with breast cancer? Ann Surg. May 1997;225(5):579-586; discussion 586-579.

[Response Ends]

2. Scientific Acceptability of Measure Properties

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

---

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

[Response Begins]

Yes

[Yes Please Explain]

There were minor changes to definitions of outcomes.

[Response Ends]

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

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

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

[Response Begins]

Definitions were updated to better conform to current clinical understanding. There are slight changes to what constitutes a morbidity event which will not affect the validity, reliability, or feasibility of the model.

[Response Ends]

sp.01. Provide the measure title.

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

[Response Begins]

Risk Adjusted Case Mix Adjusted Elderly Surgery Outcomes Measure

[Response Ends]

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

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

[Response Begins]

This is a hospital based, risk adjusted, case mix adjusted elderly surgery aggregate clinical outcomes measure of adults 65 years of age and older.

[Response Ends]

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

Please refrain from selecting the following answer option(s). We are in the process of phasing out these answer options and request that you instead select one of the other answer options as they apply to your measure.

Please do not select:

- *Surgery: General*

[Response Begins]

Surgery

[Response Ends]

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

[Response Begins]

Safety: Complications

[Response Ends]

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

Select only those target populations which can be stratified in the reporting of the measure's result.

Please refrain from selecting the following answer option(s). We are in the process of phasing out these answer options and request that you instead select one of the other answer options as they apply to your measure.

Please do not select:

- *Populations at Risk: Populations at Risk*

[Response Begins]

Elderly (Age >= 65)

[Response Ends]

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

Check ONLY the levels of analysis for which the measure is SPECIFIED and TESTED.

Please refrain from selecting the following answer option(s). We are in the process of phasing out these answer options and request that you instead select one of the other answer options as they apply to your measure.

Please do not select:

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

[Response Begins]

Facility

[Response Ends]

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

Check ONLY the settings for which the measure is SPECIFIED and TESTED.

[Response Begins]

Inpatient/Hospital

Outpatient Services

[Response Ends]

sp.09. Provide a URL link to a web page specific for this measure that contains current detailed specifications including code lists, risk model details, and supplemental materials.

Do not enter a URL linking to a home page or to general information. If no URL is available, indicate "none available".

[Response Begins]

None available

[Response Ends]

sp.12. Attach the data dictionary, code table, or value sets (and risk model codes and coefficients when applicable). Excel formats (.xlsx or .csv) are preferred.

Attach an excel or csv file; if this poses an issue, contact staff. Provide descriptors for any codes. Use one file with multiple worksheets, if needed.

[Response Begins]

No data dictionary/code table – all information provided in the submission form

[Response Ends]

For the question below: state the outcome being measured. Calculation of the risk-adjusted outcome should be described in sp.22.

sp.13. State the numerator.

Brief, narrative description of the measure focus or what is being measured about the target population, i.e., cases from the target population with the target process, condition, event, or outcome).

DO NOT include the rationale for the measure.

[Response Begins]

2022 Maintenance

This measure examines the occurrence of a mortality/morbidity composite defined later in the submission. NSQIP routinely adjusts definitions of outcomes for reasons that could include: Enhancement of clinical meaningfulness, simplicity, concordance with definitions constructed by other entities, and so forth. For the most part, these changes are minor and have limited impact on which cases are defined as having experienced an event or not. Furthermore, as revised definitions apply to all patients at participating hospitals, these changes do not impact the fairness of risk adjustment.

2022 definitions are described in section sp13 in reference to 2016 definitions.

2017 Maintenance

The outcome of interest is hospital-specific risk-adjusted mortality, a return to the operating room, or any of the following morbidities as defined by American College of Surgeons National Surgical Quality Improvement Program (ACS NSQIP): Cardiac Arrest requiring CPR, Myocardial Infarction, Sepsis, Septic Shock, Deep Incisional Surgical Site Infection (SSI), Organ/Space SSI, Wound Disruption, Unplanned Reintubation without prior ventilator dependence, Pneumonia without pre-operative pneumonia, progressive Renal Insufficiency or Acute Renal Failure without pre-operative renal failure or dialysis, or urinary tract infection (UTI) within 30 days of any ACS NSQIP listed (CPT) surgical procedure. The original endorsed measure included venous thromboembolism (VTE) as eligible morbidity events, including deep venous thrombosis requiring therapy and pulmonary embolism.

[Response Ends]

For the question below: describe how the observed outcome is identified/counted. Calculation of the risk-adjusted outcome should be described in sp.22.

sp.14. Provide details needed to calculate the numerator.

All information required to identify and calculate the cases from the target population with the target process, condition, event, or outcome such as definitions, time period for data collection, specific data collection items/responses, code/value sets.

Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at sp.11.

[Response Begins]

2022 Maintenance

2017 definitions are unbolded; comments and changes reflecting revised definitions are **bolded**.

Mortality- "All cause" Death within the 30-day follow-up period: Any death occurring through midnight on the 30th day after the date of the procedure, regardless of cause, in or out of the hospital.

No Change to mortality

All other outcome fields also defined explicitly in the tradition of ACS NSQIP:

Unplanned reoperation: Patient had an unplanned return to the operating room for a surgical procedure related to either the index or concurrent procedure performed. This return must be within the 30 day postoperative period. The return to the OR may occur at any hospital or surgical facility (i.e. your hospital or at an outside hospital).

An unplanned reoperation no longer has to be related to the surgery; any unplanned reoperation qualifies. A determination was made that "relateness" was too subjective and led to inconsistent assignments across hospitals.

Cardiac Arrest Requiring CPR: The absence of cardiac rhythm or presence of chaotic cardiac rhythm that results in loss of consciousness requiring the initiation of any component of basic and/or advanced cardiac life support. Patients with automatic implantable cardioverter defibrillator (AICD) that fire but the patient has no loss of consciousness should be excluded.

The absence of cardiac rhythm or presence of a cardiac rhythm requiring the initiation of cardiopulmonary resuscitation. Criteria: The following criteria, A, B or C below, must be noted intraoperatively or within 30 days after the primary procedure: A. The absence of a cardiac rhythm or presence of cardiac rhythm requiring the initiation of chest compressions OR B. Patients in V-Fib or pulseless VT in which defibrillation is performed with

or without chest compressions OR C. Patients with automatic implantable cardioverter defibrillator (AICD) that fires and the patient has loss of consciousness.

This definition was changed to provide more detailed assignment rules.

Myocardial Infarction: An acute myocardial infarction occurring within 30 days following surgery as manifested by one of the following three criteria:

a. Documentation of ECG changes indicative of acute MI (one or more of the following):

• ST elevation > 1 mm in two or more contiguous leads

• New left bundle branch

• New q-wave in two of more contiguous leads

b. New elevation in troponin (**including high sensitivity tropinin assays)** including high sensitivity tropinin assays) greater than 3 times upper level of the reference range in the setting of suspected myocardial ischemia

c. Physician diagnosis of myocardial infarction.

Updated because more sites are using this assay.

Sepsis: Sepsis is the systemic response to infection. Report this variable if the patient has TWO OR MORE of the following five clinical signs and symptoms of Systemic Inflammatory Response Syndrome (SIRS):

a. Temp >38 degrees C (100.4 degrees F) or < 36 degrees C (96.8 degrees F)

b. HR >90 bpm

c. RR >20 breaths/min or PaCO2 <32 mmHg(<4.3 kPa)

d. WBC >12,000 cell/mm3, <4000 cells/mm3, or >10% immature (band) forms

e. Anion gap acidosis: this is defined by either:

• [Na + K] – [Cl + HCO3 (or serum CO2)]. If this number is greater than 16, then an anion gap acidosis is present.

• Na – [Cl + HCO3 (or serum CO2)]. If this number is greater than 12, then an anion gap acidosis is present.

AND one of the following:

a. positive blood culture

b. clinical documentation of purulence or positive culture from any site thought to be causative

In addition, a patient with a suspected post-operative clinical condition of infection, or bowel infarction, (which leads to the surgical procedure and meets the criteria for SIRS above), the findings at operation must confirm the diagnosis with one of more of the following:

• Confirmed infarcted bowel requiring resection

• Purulence in the operative site

• Enteric contents in the operative site, or

• Positive intra-operative cultures

Severe Sepsis/Septic Shock: Sepsis is considered severe when it is associated with organ and/or circulatory dysfunction. Report this variable if the patient has sepsis AND documented organ and/or circulatory dysfunction. Examples of organ dysfunction include: oliguria, acute alteration in mental status, acute respiratory distress. Examples of circulatory dysfunction include: hypotension, requirement of inotropic or vasopressor agents. Severe Sepsis/Septic Shock is assigned when it appears to be related to Sepsis and not a Cardiogenic or Hypovolemic etiology.

No change to Sepsis

Deep Incisional SSI: Deep Incision SSI is an infection that occurs within 30 days after the operation and the infection appears to be related to the operation and infection involved deep soft tissues (for example, fascial and muscle layers) of the incision and at least one of the following:

• Purulent drainage **or a positive culture** from the deep incision but not from the organ/space component of the surgical site.

• A deep incision spontaneously dehisces or is deliberately opened by a surgeon when the patient has at least one of the following signs or symptoms: fever (> 38 C), localized pain, or tenderness, unless site is culture-negative.

• An abscess or other evidence of infection involving the deep incision is found on direct examination, during reoperation, or by histopathologic or radiologic examination.

• Diagnosis of a deep incision SSI by a surgeon or attending physician.

Last 2 items changed to

An abscess or other evidence of infection involving the deep incision is found on direct examination, during reoperation, or radiologic examination.

•Diagnosis of a deep incision SSI by a physician or advanced provider.

Organ/Space SSI: is an infection that occurs within 30 days after the operation and the infection appears to be related to the operation and the infection involves any part of the anatomy (for example, organs or spaces), other than the incision, which was opened or manipulated during an operation and at least one of the following:

• Purulent drainage from a drain that is placed through a stab wound into the organ/space.

• Organisms isolated from an aseptically obtained culture of fluid or tissue in the organ/space.

• An abscess or other evidence of infection involving the organ/space that is found on direct examination, during reoperation, or by histopathologic or radiologic examination.

• Diagnosis of an organ/space SSI by a surgeon or attending physician.

Last 2 items changed to

An abscess or other evidence of infection involving the deep incision is found on direct examination, during reoperation, or radiologic examination.

Diagnosis of a deep incision SSI by a physician or advanced provider.

Wound Disruption: Separation of the layers of a surgical wound, which may be partial or complete, with disruption of the fascia.

A spontaneous reopening of a surgically closed wound that occurs within 30 days after the primary procedure AND one of the following criteria A OR B below: A. Abdominal site: • A loss of the integrity of fascial closure (or whatever closure was performed in the absence of fascial closure) OR B. Other Surgical Sites: • A spontaneous disruption or dehiscence of all layers of the surgical wound OR • A spontaneous disruption or dehiscence of part of the surgical wound that requires intervention for closure

Updated to clarify documentation of disruption on sites other than abdominal sites without fascia.

Unplanned Intubation for Respiratory/Cardiac Failure: Patient required placement of an endotracheal tube and mechanical or assisted ventilation because of the onset of respiratory or cardiac failure manifested by severe respiratory distress, hypoxia, hypercarbia, or respiratory acidosis. In patients who were intubated for their surgery, unplanned intubation occurs after they have been extubated after surgery. In patients who were not intubated during surgery, intubation at any time after their surgery is considered unplanned.

New definition:

The variable intent is to capture all unplanned intubations for any reason/cause, including, but not limited to, unplanned intubations for refractory hypotension, cardiac arrest, and inability to protect airway. Definition: The placement of an endotracheal tube or other similar breathing tube [Laryngeal Mask Airway (LMA), nasotracheal tube, etc.] and ventilator support. An unplanned intubation must be noted intraoperatively or within 30 days after the primary procedure.

Pneumonia (without preoperative pneumonia): Enter “Yes” if the patient has pneumonia meeting the definition below. Patients with pneumonia must meet criteria from both Radiology and Signs/Symptoms/Laboratory sections listed as follows:

Radiology:

One definitive chest radiological exam (x-ray or CT)* with at least one of the following:

• New or progressive and persistent infiltrate

• Consolidation or opacity

• Cavitation

*Note: In patients with underlying pulmonary or cardiac disease (e.g. respiratory distress syndrome, bronchopulmonary dysplasia, pulmonary edema, or chronic obstructive pulmonary disease), two or more serial chest radiological exams (x-ray or CT) are required. (Serial radiological exams should be taken no less than 12 hours apart, but not more than 7 days apart. The occurrence should be assigned on the date the patient first met all of the criteria of the definition i.e, if the patient meets all PNA criteria on the day of the first xray, assign this date to the occurrence. Do not assign the date of the occurrence to when the second serial xray was performed).

Signs/Symptoms/Laboratory:

FOR ANY PATIENT, at least one of the following:

• Fever (>380C or >100.40F) with no other recognized cause

• Leukopenia (<4000 WBC/mm3) or leukocytosis(=12,000 WBC/mm3)

• For adults = 70 years old, altered mental status with no other recognized cause

And

At least one of the following:

• 5% Bronchoalveolar lavage (BAL) -obtained cells contain intracellular bacteria on direct microscopic exam (e.g., Gram stain)

• Positive growth in blood culture not related to another source of infection

• Positive growth in culture of pleural fluid

• Positive quantitative culture **or corresponding semi-quantitative culture** from minimally contaminated lower respiratory tract (LRT) specimen (e.g. BAL or protected specimen brushing)

Updated to be more inclusive of tests utilized by sites

OR

At least two of the following:

• New onset of purulent sputum, or change in character of sputum, or increased respiratory secretions, or increased suctioning requirements

• New onset or worsening cough, or dyspnea, or tachypnea

• Rales or rhonchi

• Worsening gas exchange (e.g. O2 desaturations (e.g., PaO2/FiO2 = 240), increased oxygen requirements, or increased ventilator demand)

Progressive Renal Insufficiency (without preoperative renal failure or dialysis): The reduced capacity of the kidney to perform its function as evidenced by a rise in creatinine of >2 mg/dl. but with no requirement for dialysis **(without PATOS - *preop dialysis, or preop stage 3, or preop creatinine > 4, if measured)***

Current: If the patient did not require preoperative (within the 2-week timeframe prior to surgery) or postoperative dialysis AND at least ONE of the following criterion points listed for must be met within 30 days after the primary procedure.

Creatinine Increase: An increase in serum creatinine based on two measurements, the latter of which must be within the 30-day postoperative timeframe.

A second creatinine value that has increased to 2 to <3 times within 7 days of the first value (this is stage 2, below are stage 3)

A second creatinine value that has increased to ≥3.0 times within 7 days of the first value

A second creatinine value that is ≥4.0 mg/dL (≥353.6 µmol/L) and has risen ≥0.3 mg/dL (≥26.5 µmol/L) within 48 hours from the first value

A second creatinine value that is ≥4.0 mg/dL (≥353.6 µmol/L) and has increased to ≥1.5 times

within 7 days from the first value

Updated to better align with the KDIGO stages of renal injury/failure

Acute Renal Failure Requiring Dialysis (without preoperative renal failure or dialysis): In a patient who did not require dialysis preoperatively, worsening of renal dysfunction postoperatively requiring hemodialysis, peritoneal dialysis, hemofiltration, hemodiafiltration, or ultrafiltration.

Urinary Tract Infection: Postoperative symptomatic urinary tract infection must meet ONE of the following TWO criteria:

Criterion One. One of the following **six [remove five]**:

a. fever (>38 degrees C),

b. urgency,

c. frequency,

d. dysuria,

e. suprapubic tenderness

f. costovertebral angle pain or tenderness

AND a urine culture of > 100,000 colonies/ml urine with no more than two species of organisms.

OR

Criterion Two. Two of the following five:

a. fever (>38 degrees C),

b. urgency,

c. frequency,

d. dysuria,

e. suprapubic tenderness

f. costovertebral angle pain or tenderness

AND ANY ONE or MORE of the following **[remove seven]**:

a, b, and c deleted

a. Dipstick test positive for leukocyte esterase and/or nitrate,

b. Pyuria (>10 WBCs/mm3 or > 3 WBC/hpf of unspun urine),

c. Organisms seen on Gram stain of unspun urine,

d. Two urine cultures with repeated isolation of the same uropathogen with >100 colonies/ml urine in non-voided specimen,

e. Urine culture with < 100,000 colonies/ml urine of single uropathogen in patient being treated with appropriate antimicrobial therapy,

f. Physician´s diagnosis,

g. Physician institutes appropriate antimicrobial therapy.

[Response Ends]

For the question below: state the target population for the outcome. Calculation of the risk-adjusted outcome should be described in sp.22.

sp.15. State the denominator.

Brief, narrative description of the target population being measured.

[Response Begins]

2017 Maintenance

Patients undergoing any ACS NSQIP listed (CPT) surgical procedure who are 65 years of age or older. (See appendix of roughly 2900 ACS NSQIP eligible CPT codes)

[Response Ends]

For the question below: describe how the target population is identified. Calculation of the risk-adjusted outcome should be described in sp.22.

sp.16. Provide details needed to calculate the denominator.

All information required to identify and calculate the target population/denominator such as definitions, time period for data collection, specific data collection items/responses, code/value sets.

Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at sp.11.

[Response Begins]

2017 Maintenance

Cases are collected so as to match ACS NSQIP inclusion and exclusion criteria, thereby permitting valid application of ACS NSQIP model-based risk adjustment.

[Response Ends]

sp.17. Describe the denominator exclusions.

Brief narrative description of exclusions from the target population.

[Response Begins]

2017 Maintenance

Cases must first have ACS NSQIP eligible CPT codes on the submitted list.

[Response Ends]

sp.18. Provide details needed to calculate the denominator exclusions.

All information required to identify and calculate exclusions from the denominator such as definitions, time period for data collection, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at sp.11.

[Response Begins]

2017 Maintenance

NOT ON ELIGIBLE CPT LIST: Approximately 2900 codes are eligible.

MAJOR TRAUMA: A patient who is admitted to the hospital with acute major or multisystem trauma and has surgery for that trauma is excluded, though any operation performed after the patient has been discharged from that trauma admission can be included. Exclusion of trauma cases does consider magnitude of injuries. If there are multiple severe injuries and the situation is emergent, the case would be excluded. If the patient has minor injuries, they are not excluded. For instance, ground level falls or low-velocity / low-impact injury mechanism may produce a single bone fracture (single system injury) and would be included. In contrast, a fall from a ladder (or a fall from height) would be excluded due to high-velocity / high-impact mechanism and the resulting injuries would be considered multisystem trauma. Any emergent, major or multisystem trauma case is excluded. These algorithms are communicated to the data collectors via educational tools.

TRANSPLANT: A patient who is admitted to the hospital for a transplant and has a transplant procedure and any additional surgical procedures during the transplant hospitalization will be excluded, tough any operation performed after the patient has been discharged from the transplant stay is eligible for selection.

ASA 6: A patient classified as ASA Class 6 is not eligible for inclusion.

[Response Ends]

sp.19. Provide all information required to stratify the measure results, if necessary.

Include the stratification variables, definitions, specific data collection items/responses, code/value sets, and the risk-model covariates and coefficients for the clinically-adjusted version of the measure when appropriate. Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format in the Data Dictionary field.

[Response Begins]

2017 Maintenance

The measure is risk adjusted and case mix adjusted.

[Response Ends]

sp.20. Is this measure adjusted for socioeconomic status (SES)?

[Response Begins]

[Response Ends]

sp.21. Select the risk adjustment type.

Select type. Provide specifications for risk stratification and/or risk models in the Scientific Acceptability section.

[Response Begins]

Statistical risk model

[Response Ends]

sp.22. Select the most relevant type of score.

Attachment: If available, please provide a sample report.

[Response Begins]

Ratio

[Response Ends]

sp.23. Select the appropriate interpretation of the measure score.

Classifies interpretation of score according to whether better quality or resource use is associated with a higher score, a lower score, a score falling within a defined interval, or a passing score

[Response Begins]

Better quality = Lower score

[Response Ends]

sp.24. Diagram or describe the calculation of the measure score as an ordered sequence of steps.

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

[Response Begins]

2022 Maintenance

The measure score is actually a risk-adjusted, smoothed (BLUP estimate via empirical Bayes) odds ratio. A stepped sequence is inappropriate for this process as the hierarchical model (SAS PROC GLIMMIX) does these processes together in the background. The end result is an odds ratio that describes the odds of the event at the target hospital compared to the odds at the NSQIP-estimated average hospital, if that hospital were to do the same procedures on the same patients.

2017 Maintenance

For data collected during the one year time interval at each hospital: (a) O = the number of observed adverse events at the hospital; (b) using parameters from the applicable model derived logistic equation, compute predicted event probabilities for each patient in the hospital's data set; (c) the sum of these predicted probabilities defines E; (d) compute the hospital's O/E ratio and applicable confidence intervals.

[Response Ends]

sp.27. If measure testing is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.

Examples of samples used for testing:

• Testing may be conducted on a sample of the accountable entities (e.g., hospital, physician). The analytic unit specified for the particular measure (e.g., physician, hospital, home health agency) determines the sampling strategy for scientific acceptability testing.

• The sample should represent the variety of entities whose performance will be measured. The 2010 Measure Testing Task Force recognized that the samples used for reliability and validity testing often have limited generalizability because measured entities volunteer to participate. Ideally, however, all types of entities whose performance will be measured should be included in reliability and validity testing.

• The sample should include adequate numbers of units of measurement and adequate numbers of patients to answer the specific reliability or validity question with the chosen statistical method.

• When possible, units of measurement and patients within units should be randomly selected.

[Response Begins]

2022 Maintenance

Hospitals participate in either the Essential program or the Essential program augmented with surgical Targets, which defines (approximately) whether cases are systematically sampled from all eligible CPT codes or, additionally, are sampled from one or more specific surgical groups. Regardless, cases are selected without regard to outcomes. Elderly patients represent a large portion of the surgical population and necessary sample sizes are generally achieved without additional sampling protocols (as NSQIP hospital typical collect 1600 cases per year).

2017 Maintenance

For each data collection year, hospitals estimate their number of qualifying surgeries. Based on that denominator and the required sample size to achieve reliability of 0.4 (estimated sample size for reliability 0.4 is approximately 115 cases, see Measure Testing), hospitals take a systematic sample (e.g., every 3rd qualifying case), to achieve the minimum sample size. In the event that the required sample size cannot be achieved, hospitals may collect data on all eligible patients.

[Response Ends]

sp.30. Select only the data sources for which the measure is specified.

[Response Begins]

Electronic Health Data

Electronic Health Records

Management Data

Paper Medical Records

Registry Data

[Response Ends]

sp.31. Identify the specific data source or data collection instrument.

For example, provide the name of the database, clinical registry, collection instrument, etc., and describe how data are collected.

[Response Begins]

2022 Maintenance

To clarify, this measure is potentially available to non-NSQIP participants, if either another entity were to provide benchmarking results or if the non-NSQIP hospital were to apply to NSQIP to be included in the model (available once annually). This model, including non-NSQIP hospitals, would be run separately from the routine reports

delivered to participants. To date, no non-NSQIP hospital has requested this service and, at this time, we cannot commit to providing this service in the future.

2017 Maintenance

The modeling presented herein is based on ACS NSQIP Data files for the last several years. As a measure, data are collected and reported on an annual basis. Hospitals are not required to participate in ACS NSQIP- they would simply submit their data to the implementing organization or agency, and would receive their assessments in return.

[Response Ends]

sp.32. Provide the data collection instrument.

[Response Begins]

No data collection instrument provided

[Response Ends]

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

Please separate added or updated information from the most recent measure evaluation within each question response in the Scientific Acceptability sections. For example:

Current Submission:

Updated testing information here.

Previous Submission:

Testing from the previous submission here.

[Response Begins]

No

[Response Ends]

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

Please separate added or updated information from the most recent measure evaluation within each question response in the Scientific Acceptability sections. For example:

Current Submission:

Updated testing information here.

Previous Submission:

Testing from the previous submission here.

[Response Begins]

No

[Response Ends]

2ma.03. For outcome, patient-reported outcome, resource use, cost, and some process measures, risk adjustment/stratification may be conducted. Did you perform a risk adjustment or stratification analysis?

[Response Begins]

Yes

[Response Ends]

2ma.04. For maintenance measures in which risk adjustment/stratification has been performed, indicate whether additional risk adjustment testing has been conducted since the most recent maintenance evaluation. This may include updates to the risk adjustment analysis with additional clinical, demographic, and social risk factors.

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

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

[Response Begins]

Yes - Additional risk adjustment analysis is included

[Response Ends]

Measure testing must demonstrate adequate reliability and validity in order to be recommended for endorsement. Testing may be conducted for data elements and/or the computed measure score. Testing information and results should be entered in the appropriate fields in the Scientific Acceptability sections of the Measure Submission Form.

• Measures must be tested for all the data sources and levels of analyses that are specified. If there is more than one set of data specifications or more than one level of analysis, contact NQF staff about how to present all the testing information in one form.

• All required sections must be completed.

• For composites with outcome and resource use measures, Questions 2b.23-2b.37 (Risk Adjustment) also must be completed.

• If specified for multiple data sources/sets of specifications (e.g., claims and EHRs), Questions 2b.11-2b.13 also must be completed.

• An appendix for supplemental materials may be submitted (see Question 1 in the Additional section), but there is no guarantee it will be reviewed.

• Contact NQF staff with any questions. Check for resources at the Submitting Standards webpage.

• For information on the most updated guidance on how to address social risk factors variables and testing in this form refer to the release notes for the 2021 Measure Evaluation Criteria and Guidance.

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

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

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

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

AND

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

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

• an evidence-based risk-adjustment strategy (e.g., risk models, risk stratification) is specified; is based on patient factors (including clinical and social risk factors) that influence the measured outcome and are present at start of care; 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 non-responders) and how the specified handling of missing data minimizes bias.

2c. For composite performance measures, empirical analyses support the composite construction approach and demonstrate that:

2c1. the component measures fit the quality construct and add value to the overall composite while achieving the related objective of parsimony to the extent possible; and

2c2. the aggregation and weighting rules are consistent with the quality construct and rationale while achieving the related objective of simplicity to the extent possible.

(if not conducted or results not adequate, justification must be submitted and accepted)

Definitions

Reliability testing applies to 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).

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.

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

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

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

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.

Please separate added or updated information from the most recent measure evaluation within each question response in the Scientific Acceptability sections. For example:

Current Submission:

Updated testing information here.

Previous (Year) Submission:

Testing from the previous submission here.

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

[Response Begins]

Registry Data

[Response Ends]

2a.02. If an existing dataset was used, identify the specific dataset.

The dataset used for testing must be consistent with the measure specifications for target population and healthcare entities being measured; e.g., Medicare Part A claims, Medicaid claims, other commercial insurance, nursing home MDS, home health OASIS, clinical registry).

[Response Begins]

Original submission

See Risk-adjustment Methodology in Specifications. Models were constructed using a large sample derived from the ACS NSQIP database for 2008.

May 31, 2016 Maintenance of Endorsement Update:

See Risk-adjustment Methodology in Specifications.

Models were constructed using a large systematic and unbiased sample from the ACS NSQIP database for July 1, 2011 through June 30, 2015 (refereed to henceforth as years 2011-2014) yielding 655,187 patient records for eligible surgeries from 509 hospitals. Evaluation of measure performance, in the context of prospective quality assessments, are based on the analysis of hospital data for one year (2014, hospitals = 460, cases = 206,064), with those data being analyzed using the historical equations derived from the 2011-2014 dataset.

[Response Ends]

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

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

[Response Begins]

01-01-2011 - 12-31-2014

[Response Ends]

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

Testing must be provided for all the levels specified and intended for measure implementation, e.g., individual clinician, hospital, health plan.

Please refrain from selecting the following answer option(s). We are in the process of phasing out these answer options and request that you instead select one of the other answer options as they apply to your measure.

Please do not select:

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

[Response Begins]

Facility

[Response Ends]

2a.05. List the measured entities included in the testing and analysis (by level of analysis and data source).

Identify the number and descriptive characteristics of measured entities included in the analysis (e.g., size, location, type); if a sample was used, describe how entities were selected for inclusion in the sample.

[Response Begins]

2017 Maintenance

655,187 patient records for eligible surgeries from 509 hospitals

[Response Ends]

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

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

[Response Begins]

2022 Maintenance

Because of the number of hospitals participating and number of patients included, the NSQIP database would be representative of the elderly surgical population with respect to age, sex, race, and indications for surgery. Values from the original submission are unavailable.

[Response Ends]

2a.07. If there are differences in the data or sample used for different aspects of testing (e.g., reliability, validity, exclusions, risk adjustment), identify how the data or sample are different for each aspect of testing.

[Response Begins]

2022 Maintenance

The same registry-based data set was used for all aspects of testing

[Response Ends]

2a.08. List the social risk factors that were available and analyzed.

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

[Response Begins]

May 31, 2016 Maintenance of Endorsement Update:

Using a 95% confidence interval for the observed to expected events (O/E) ratio, the original elderly surgery outcome measure (without SES and with VTE) identified 49 low and 34 high outliers among the 460 hospitals with data in 2014. The addition of SES data changed the outlier status among few hospitals: 3 high outliers shifted to no outlier status, 3 low outliers shifted to no outlier status, and 3 hospitals previously not outliers became high (n=2) or low (n=1) outlier status. (weighted kappa = 0.9287). In addition, 21 hospitals increased decile status by 1 category and 21 hospitals decreased decile status by 1 category. These data suggest that SES-related variables are not influential in risk adjustment with respect to the 30-day elderly surgery outcome measure.

We also examined the effect of removing VTE for models without SES variables. The comparison of outlier determinations is shown below (weighted kappa = 0.5982)

Outlier Status (n)		Elderly surgery measure, with VTE without SES			
		HIGH	LOW	NO	Total
Elderly surgery measure, without VTE without SES	HIGH	33	0	31	64
	LOW	0	22	0	22
	NO	1	27	346	374
	Total	34	49	377	460

There were several changes in decile status with the removal of VTE:

Difference: Decile without VTE - decile with VTE

Decile difference	Number of Hospitals
-2	9
-1	75
0	290
1	79
2	7

The inclusion versus exclusion of VTE has important effects on outlier and decile status.

[Response Ends]

Note: If accuracy/correctness (validity) of data elements was empirically tested, separate reliability testing of data elements is not required – in 2a.09 check patient or encounter-level data; in 2a.010 enter "see validity testing section of data elements"; and enter "N/A" for 2a.11 and 2a.12.

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

Choose one or both levels.

[Response Begins]

Patient or Encounter-Level (e.g., inter-abstractor reliability; data element reliability must address ALL critical data elements)

[Response Ends]

2a.10. For each level of reliability testing checked above, describe the method of reliability testing and what it tests.

Describe the steps—do not just name a method; what type of error does it test; what statistical analysis was used.

[Response Begins]

2a2.1 Data/Sample *(Description of the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)***:**

See Risk-adjustment Methodology in Specifications. Models were constructed using a large sample derived from the ACS NSQIP database for 2008.

May 31, 2016 Maintenance of Endorsement Update:

See Risk-adjustment Methodology in Specifications.

Models were constructed using a large systematic and unbiased sample from the ACS NSQIP database for July 1, 2011 through June 30, 2015 (refereed to henceforth as years 2011-2014) yielding 655,187 patient records for eligible surgeries from 509 hospitals. Evaluation of measure performance, in the context of prospective quality assessments, are based on the analysis of hospital data for one year (2014, hospitals = 460, cases = 206,064), with those data being analyzed using the historical equations derived from the 2011-2014 dataset.

2a2.2 Analytic Method *(Describe method of reliability testing & rationale)***:**

See Risk-adjustment Methodology in Specifications. Reliability was determined using ICCs estimated by SAS PROC GENMOD.

May 31, 2016 Maintenance of Endorsement Update:

Reliability was assessed using a standard method (described in: Huffman, K.M., Cohen, M.E, Ko, C.Y., Hall, B.L. A comprehensive evaluation of statistical reliability in ACS NSQIP profiling models. Annals of Surgery, 2015, 261, 1108-1113), which uses information provided by a random intercept, fixed slope, hierarchical model (implemented by SAS PROC GLIMMIX).

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

See Risk-adjustment Methodology in Specifications. The relative variation between hospitals defined by the intra-class correlation coefficient (ICC) for hospitals can be estimated for continuous outcomes using linear mixed models, but the within-hospital variation needed to calculate ICCs is not routinely estimated for dichotomous outcomes. Hence, the usual measure of ICC based on a latent variable formulation using the standard logistic distribution was estimated. The between-hospital variation component of the ICC was estimated from SAS PROC GENMOD regressing the composite outcome on the significant predictors for mortality/serious morbidity in patients >=65. Together with procedure volumes, these ICCs were entered into the following equation to estimate reliability:

$R = nICC/(1 + (n -1)ICC)$,

where R is the reliability, n is the case load per hospital and ICC is the intra-class correlation.

There are no definitive criteria for what level of reliability is acceptable, but it is proposed to be similar to inter-rater reliability standards used for assessing survey instruments.

RELIABILITY ESTIMATE	INTERPRETATION
0.00-0.20	Slight
0.21-0.40	Fair
0.41-0.60	Moderate
0.61-0.80	Substantial
0.81-1.00	Excellent

The ICC was estimated at 0.00377. Using a minimum acceptable reliability for mortality/serious morbidity in patients >=65 of 0.4(moderate), requiring roughly 180 cases, the proportions of hospitals likely to have these "moderate" reliability estimate are as follows. 90.8% of all U.S. hospitals and 84.8% of ACS NSQIP hospitals meet the 0.4 reliability requirement. It is estimated that >95% of all eligible cases performed in the country would be captured within this institutional set.

Table 1. Estimates of Procedure Volume Required to Achieve Specified Measure Reliability, and Proportions of U.S. Hospitals and ACS NSQIP Hospitals Meeting the Volume Requirements.

Reliability__RequiredCases__%U..S.HospMtgRqrmnt*__%NSQIPHospMtg Rqrmnt+

_____0.3_____114_____93.3_____92.4

_____0.4_____177_____90.8_____84.8

_____0.5_____265_____86.8_____72.5

_____0.6_____397_____80.9_____46.5

_____0.7_____617_____70.7_____13.3

*Based on volume data from the 2005 National Inpatient Survey and inflated to account for outpatient procedures.

+Based on ACS NSQIP Data file 2008 and inflated to account for procedures that might be excluded for over-representation

May 31, 2016 Maintenance of Endorsement Update:

For Measure reliability (understood here as the ability to differentiate quality between hospitals) in the context of data collected during a single year, we evaluated reliability for 460 hospitals collecting data during 2014. As described in sections 2b2.2 and 2b4.2, we are also interested in evaluating the effects of 2 separate adjustments to the Elderly surgery outcome measure: (1) dropping venous thromboembolic (VTE) events as a component of the outcome; and (2) inclusion of socioeconomic status (SES)-related variables for risk adjustment. Therefore, reliability in the 2014 dataset was examined under the 4 conditions defined by the factorial combination of VTE (included or not included) and SES variables (included or not included). The table describes the percentage of hospitals for which the measure provides the indicated level of statistical reliability for hospitals providing data in 2014.

Calibration range	Percent: with VTE, without SES (original model)	Percent: with VTE, with SES	Percent: without VTE, without SES	Percent: without VTE, with SES
0.00-0.20	11.09	11.09	10.43	10.43
0.21-0.40	5.65	6.09	5.65	5.65
0.41-0.60	15.22	14.78	14.13	14.35
0.61-0.80	49.35	49.35	46.96	47.17
0.81-1.00	18.70	18.70	22.83	22.39

Using a minimum acceptable reliability of 0.4, the proportions of hospitals with a “minimally acceptable” reliability estimate for the elderly surgery outcome measure is excellent, totaling above 80% across all four variations.

The mean number cases per hospital in the 2014 data set was 448, but there was positive skew in the distribution of sample sizes (median=437). We generated a nonlinear regression equation, predicting hospital reliability from hospital sample size using the model that eliminated VTE and did not include SES variables (this is the approach that will be recommended for this measure). It must be understood that reliability is dependent on several factors, but most notably sample size and the true magnitude of hospital quality differences. The regression plot, estimated from this dataset, is shown below.

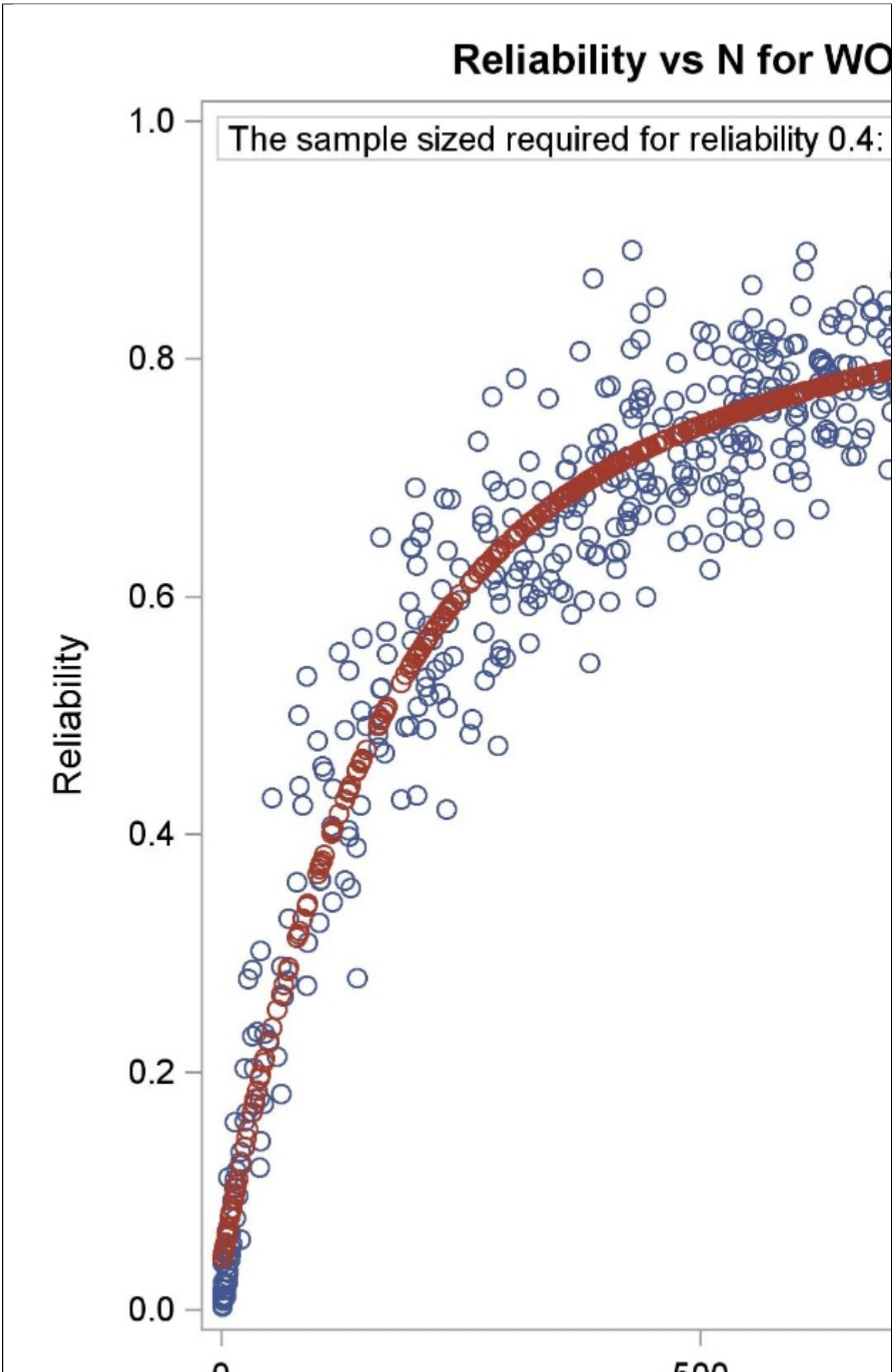
Reliability vs N for WO
The sample sized required for reliability 0.4:
Reliability
1.0
0.8
0.6
0.4
0.2
0.0

Figure: Reliability vs Case number for Elderly Surgery measure (without VTE and without SES), regression and observed

[Response Ends]

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

For example, provide the percent agreement and kappa for the critical data elements, or distribution of reliability statistics from a signal-to-noise analysis. For score-level reliability testing, when using a signal-to-noise analysis, more than just one overall statistic should be reported (i.e., to demonstrate variation in reliability across providers). If a particular method yields only one statistic, this should be explained. In addition, reporting of results stratified by sample size is preferred (pg. 18, NQF Measure Evaluation Criteria).

[Response Begins]

Thus, an empirical estimate of the sample size required to achieve reliability of 0.4 is 115. This number is considered an appropriate and achievable target for the number of cases for hospitals interested in participating in this measure given current case numbers. The majority of hospitals (>75%) currently submit sufficient cases for high reliability. Nevertheless, hospitals with fewer cases might still benefit from participation in this measure.

[Response Ends]

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

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

[Response Begins]

2017 Maintenance

This number [115] is considered an appropriate and achievable target for the number of cases for hospitals interested in participating in this measure given current case numbers. The majority of hospitals (>75%) currently submit sufficient cases for high reliability. Nevertheless, hospitals with fewer cases might still benefit from participation in this measure.

[Response Ends]

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

[Response Begins]

Accountable Entity Level (e.g. hospitals, clinicians)

Empirical validity testing

[Response Ends]

2b.02. For each level of testing checked above, describe the method of validity testing and what it tests.

Describe the steps—do not just name a method; what was tested, e.g., accuracy of data elements compared to authoritative source, relationship to another measure as expected; what statistical analysis was used.

[Response Begins]

2017 Maintenance

Data element validity for the ACS NSQIP has been established through rigorous definitions, consistent training for data abstracters, community exchanges, and regular formal data audits. With over 600 participating hospitals, data audits are performed on 5-10% of participating sites annually. This equates to audit of over 12,600 data elements annually with extremely low disagreement rates. During the 2015 audits, the average disagreement rate was between 2-3%. Prior published evaluation of ACS NSQIP inter-rater reliability have also demonstrated low rates of disagreement, 3.15% in 2005 down to 1.56% in 2008. (Shiloach et al).

According to quality models by Donabedian, Porter and others, outcomes are the endpoint of good quality care. Compared to process or infrastructure elements, outcomes have inherent face validity. The Steering Committees for each of the ACS Quality Programs, such as the ACS Geriatric Surgery Taskforce and Coalition for Quality in Geriatric Surgery, consistently consider death and morbidity events top priority outcomes for measurement and tracking. More importantly, the current elderly surgery outcome measure has been used by an estimated 30,000 surgeons contributing cases to the ACS NSQIP database over the past 11 years of its use. Ongoing feedback from and use of this measure by these "on-the-ground" surgeons indicate shared face validity of the measure.

Additional analyses described herein pertain to model validity. For the proposed evaluation of cross validation, c-statistics (discrimination), Brier score (combined discrimination and calibration), and Hosmer-Lemeshow (calibration) p-values were computed for the 2011-2014 dataset, an entirely separate dataset (2010, identified with "2010" in the column heading in the first table of 2b2.3), and for each year 2011 through 2014 (it is understood that these are not perfect assessments of cross validation as there is an approximate 25% data overlap with respect to the model-generating dataset). Different years were examined in order to evaluate degradation of model quality due to time period effects. Statistics are broken down for VTE+ and VTE- (as an eligible event for the elderly surgery measure), and for with and without SES variables, in order to assess their effects on model quality with respect to discrimination and calibration.

Shiloach M, Frencher SK, Jr., Steeger JE, et al. Toward robust information: data quality and inter-rater reliability in the American College of Surgeons National Surgical Quality Improvement Program. J Am Coll Surg. 2010 Jan;210(1):6-16.

[Response Ends]

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

Examples may include correlations or t-test results.

[Response Begins]

2017 Maintenance

In general: (a) model quality remains consistent when the 2011-2014 equations are applied to a unique dataset (2010) and when applied to subsets of data with an approximate 25% overlap; and (b) model quality is essentially unaffected by the presence versus absence of SES in the prediction equation, while removal of VTE from the outcome appears to slightly decrease the fit with an increased discrimination. Because of the very large sample sizes studied here, a statistically significant Hosmer-Lemeshow statistic is not considered informative with respect to calibration.

Model	Model c statistic	2010 c statistic	Model HL	Model p-value	2010 HL	2010 p-value	Model Brier	2010 Brier
Elderly measure With VTE, Without SES	0.7586	0.7656	30.4886	0.0002	126.5938	0.0000	0.0888	0.1008
Elderly measure With VTE, With SES	0.7588	0.7659	31.1788	0.0001	131.9001	0.0000	0.0888	0.1008
Elderly measure Without VTE, Without SES	0.7714	0.7875	68.3632	0.0000	32.8440	0.0001	0.0771	0.0825
Elderly measure Without VTE, With SES	0.7715	0.7877	66.0084	0.0000	36.3637	0.0000	0.0771	0.0825

Jul 1, 2014 - Jun 30, 2015	C Statistic	HL	p_value	Brier
Elderly measure with VTE, without SES	0.7579	18.6091	0.0171	0.0872
Elderly measure with VTE, with SES	0.7580	16.9361	0.0308	0.0872
Elderly measure without VTE, without SES	0.7637	158.9828	0.0000	0.0819
Elderly measure without VTE, with SES	0.7638	162.3122	0.0000	0.0819

Jul 1, 2013 - Jun 30, 2014	C Statistic	HL	p_value	Brier
Elderly measure with VTE, without SES	0.7559	15.9363	0.0433	0.0873
Elderly measure with VTE, with SES	0.7561	16.6497	0.0340	0.0872
Elderly measure without VTE, without SES	0.7631	176.8872	0.0000	0.0820
Elderly measure without VTE, with SES	0.7633	183.0287	0.0000	0.0820

Jul 1, 2012 - Jun 30, 2013	C Statistic	HL	p_value	Brier
Elderly measure with VTE, without SES	0.7602	24.2472	0.0021	0.0877
Elderly measure with VTE, with SES	0.7605	25.3237	0.0014	0.0877
Elderly measure without VTE, without SES	0.7865	430.8442	0.0000	0.0671
Elderly measure without VTE, with SES	0.7867	429.6263	0.0000	0.0671

Jul 1, 2011 - Jun 30, 2012	C Statistic	HL	p_value	Brier
Elderly measure with VTE, without SES	0.7599	37.7755	0.0000	0.0951
Elderly measure with VTE, with SES	0.7601	37.6694	0.0000	0.0951
Elderly measure without VTE, without SES	0.7855	195.9585	0.0000	0.0740
Elderly measure without VTE, with SES	0.7857	188.3707	0.0000	0.0740

[Response Ends]

2b.04. Provide your interpretation of the results in terms of demonstrating validity. (i.e., what do the results mean and what are the norms for the test conducted?)

[Response Begins]

2022 Maintenance

The elderly measure model has discrimination and calibration, appropriate for quality benchmarking of hospitals

[Response Ends]

2b.05. Describe the method for determining if statistically significant and clinically/practically meaningful differences in performance measure scores among the measured entities can be identified.

Describe the steps—do not just name a method; what statistical analysis was used? Do not just repeat the information provided in Importance to Measure and Report: Gap in Care/Disparities.

[Response Begins]

2022 Maintenance

The measure is part of the ongoing NSQIP set of quality assessments. Reporting of meaningful differences is represented by the rate at which hospitals are identified as statistical outliers and/or categorized as "Needs Improvement" or "Exemplary", where the latter determinations are made when either the hospital is a statistical outlier or identified in being in the 4th or 1st adjusted (smoothed) rank quartiles of odds ratios. As the adjusted quartile criteria is somewhat less difficult to achieve than being a statistical outlier, it can serve as an early warning in the absence of statistical significance.

[Response Ends]

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

Examples may include number and percentage of entities with scores that were statistically significantly different from mean or some benchmark, different from expected; how was meaningful difference defined.

[Response Begins]

2022 Maintenance

A recent NSQIP benchmarking report (January 2022 for data collected between 7/1/2020 and 6/30/2021 showed that the measure is capable of identifying a reasonable number of low and high performing hospitals.

Model (January 2022 SAR)	Sites Included	Total Cases	Observed Events	Observed Rate	Low Outliers	High Outliers	First Adjusted Quartile	Fourth Adjusted Quartile	Exemplary	Needs Improvement
MEASURE Elderly DSM (2)	686	359425	33525	9.33	56	76	98	108	99	108

[Response Ends]

2b.07. Provide your interpretation of the results in terms of demonstrating the ability to identify statistically significant and/or clinically/practically meaningful differences in performance across measured entities.

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

[Response Begins]

2022 Maintenance

In the report presented in 2b.06 8.2% of hospitals were identified as low outliers (good performance) and 11.1% as high outliers. Using the less conservative criterion influenced by smoothed quartile as well as outlier status, 14.4% of hospitals were classified as "Exemplary" and 15.7% were classified as "Needs Improvement".

[Response Ends]

2b.08. Describe the method of testing conducted to identify the extent and distribution of missing data (or non-response) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and non-responders). Include how the specified handling of missing data minimizes bias.

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

[Response Begins]

2022 Maintenance

Missing/unknown variable values are captured in the registry. Variables that comprise the Death and Morbidity composite outcome are not permitted to be missing/unknown by the registry (the case cannot be marked a complete if data are missing and will not enter the registry). Missing/unknown predictor variables are extremely rare and values are imputed. Tests have not been run to evaluate bias due to missingness due to extremely low missingness event rates (see 2b.09).

[Response Ends]

2b.09. Provide the overall frequency of missing data, the distribution of missing data across providers, and the results from testing related to missing data.

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

[Response Begins]

2022 Maintenance

Rates of missingness in a recent NSQIP dataset were:

Outcome variables used for the composite death/morbidity outcome: No missing data

Predictor variables used for risk adjustment in the elderly measure:

1. Principal CPT code: No missing data
2. ASA Class: 0.27% missing
3. Functional Status: No missing data

[Response Ends]

2b.10. Provide your interpretation of the results, in terms of demonstrating that performance results are not biased due to systematic missing data (or differences between responders and non-responders), and how the specified handling of missing data minimizes bias.

In other words, what do the results mean in terms of supporting the selected approach for missing data and what are the norms for the test conducted; if no empirical analysis was conducted, justify the selected approach for missing data.

[Response Begins]

2022 Maintenance

Missingness rates are too low to bias results.

[Response Ends]

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

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

[Response Begins]

No, there is only one set of specifications for this measure

[Response Ends]

2b.12. Describe the method of testing conducted to compare performance scores for the same entities across the different data sources/specifications.

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

[Response Begins]

[Response Ends]

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

Examples may include correlation, and/or rank order.

[Response Begins]

[Response Ends]

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

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

[Response Begins]

[Response Ends]

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

[Response Begins]

N/A or no exclusions

[Response Ends]

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

Describe the steps—do not just name a method; what was tested, e.g., whether exclusions affect overall performance scores; what statistical analysis was used?

[Response Begins]

2022 Maintenance

N/A

[Response Ends]

2b.17. Provide the statistical results from testing exclusions.

Include overall number and percentage of individuals excluded, frequency distribution of exclusions across measured entities, and impact on performance measure scores.

[Response Begins]

2022 Maintenance

N/A

[Response Ends]

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

In other words, the value outweighs the burden of increased data collection and analysis. Note: If patient preference is an exclusion, the measure must be specified so that the effect on the performance score is transparent, e.g., scores with and without exclusion.

[Response Begins]

2022 Maintenance

N/A

[Response Ends]

2b.19. Check all methods used to address risk factors.

[Response Begins]

Statistical risk model with risk factors (specify number of risk factors)

[Statistical risk model with risk factors (specify number of risk factors) Please Explain]

The statistical model uses 3 pre-operative risk adjustment variables:

1. Principal CPT code
2. ASA Class
3. Functional Status

[Response Ends]

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

[Response Begins]

2022 Maintenance

We continue to use hierarchical logistic regression with the quality metric being defined as the hospital odds ratio (the random effect). The measure is reported quarterly with parameters estimated using a contemporaneous 12-months of data (data for successive reports overlap by about 75% percent). Because of this routine reanalysis, we do not provide intercepts and parameter values as they are continually changing.

Definitions of risk-adjustment categories are as follows. Also provided are the odds ratios calculated for a recent report (January 2022 for data collected between 7/1/2020 and 6/30/2021).

Effect	Odds Ratio
Intercept	
ASA Class (2 vs. 1)	1.345(1.096, 1.651)
ASA Class (3 vs. 1)	2.260(1.844, 2.771)
ASA Class (4-5 vs. 1)	5.025(4.093, 6.169)
Functional Status (Partially Dependent vs. Independent)	1.745(1.672, 1.822)
Functional Status (Totally Dependent vs. Independent)	1.964(1.790, 2.155)
Outcome-Specific CPT Risk	2.847(2.797, 2.897)

[Response Ends]

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

[Response Begins]

[Response Ends]

2b.22. Select all applicable resources and methods used to develop the conceptual model of how social risk impacts this outcome.

[Response Begins]

Published literature

Internal data analysis

[Response Ends]

2b.23. Describe the conceptual and statistical methods and criteria used to test and select patient-level risk factors (e.g., clinical factors, social risk factors) used in the statistical risk model or for stratification by risk.

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

[Response Begins]

2022 Maintenance

Historically, we evaluate improvements in discrimination (c-statistic) and calibration (Hosmer-Lemeshow) as potential predictor variables are added to a hierarchical logistic equation. All predictor variables are preoperate and, all things being equal, variables for this measure would be those most easily defined and accessible by clinical reviewers. We have determined that this parsimonious predictor set is capable of providing useful results for these benchmarking purposes.

[Response Ends]

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

[Response Begins]

2022 Maintenance

Based on years of experience and preliminary testing done for the initial submission, the predictors selected are very efficient risk adjustment variables.

Original submission

A parsimonious predictor set was constructed from the full step-wise set. Step-wise logistic regression (P<0.05 for inclusion), which selected from a total of 26 predictors, identified 21 predictors for inclusion in the model. In order of inclusion these variables were: CPT Risk, pre-operative Functional Status, ASA Class, Emergent, history of COPD, Wound Class, Ventilator Dependent, Weight Loss, Dyspnea, Steroid Use, Disseminated Cancer, Age Group, Ascites, Smoking, Bleeding Disorder, Radio Therapy, BMI Class, Previous Vascular Event/Disease, Alcohol Use, Previous Neurological Event/Disease, and Diabetes. The c-statistic was 0.774 and the Hosmer-Lemeshow was 0.002. Because of the very large sample sizes studied here, a statistically significant Hosmer-Lemeshow statistic is not considered informative with respect to calibration. Using only the first three selected variables (Log Odds CPT Group, Functional Status, and ASA Class), which is being advocated as the risk-adjustment model, the c-statistic was 0.764 and the Hosmer-Lemeshow was 0.002. The use of these three predictors for modeling was further

evaluated. Using a 95% confidence interval for the ratio of observed to expected events (O/E), this three variable logistic model identified 30 statistical outliers (16 low outliers and 14 high outliers). When the same three-variables were used in a random intercept, fixed slope, hierarchical model (SAS PROC GLIMMIX) using only the fixed portion of the prediction equation (NOBLUP option), 28 outliers were detected (14 low outliers and 14 high outliers). Thus, using a 95% confidence interval, logistic and hierarchical models identified 7% of hospitals as high outliers.

[Response Ends]

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

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

[Response Begins]

2022 Maintenance

We do not use social risk factors in our risk adjustment for this measure for two reasons. (1) Internal analyses have demonstrated that after risk adjustment (which accounts for comorbidity burden), SES doesn't have important predictive value. (2) We hesitate to risk adjust for SES as that would, in theory, compensate hospitals for poorer performance when treating low SES or minority patients (though we don't actually see an SES or race effect) - we prefer to not to risk adjust away any challenges associated with the SES of patients. Results from work on SES done for the 2017 Maintenance are reported earlier in this document.

[Response Ends]

2b.26. Describe the method of testing/analysis used to develop and validate the adequacy of the statistical model or stratification approach (describe the steps—do not just name a method; what statistical analysis was used). Provide the statistical results from testing the approach to control for differences in patient characteristics (i.e., case mix) below. If stratified ONLY, enter "N/A" for questions about the statistical risk model discrimination and calibration statistics.

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

[Response Begins]

2017 Measure Maintenance

Additional analyses described herein pertain to model validity. For the proposed evaluation of cross validation, c-statistics (discrimination), Brier score (combined discrimination and calibration), and Hosmer-Lemeshow (calibration) p-values were computed for the 2011-2014 dataset, an entirely separate dataset (2010, identified with "2010" in the column heading in the first table of 2b2.3), and for each year 2011 through 2014 (it is understood that these are not perfect assessments of cross validation as there is an approximate 25% data overlap with respect to the model-generating dataset). Different years were examined in order to evaluate degradation of model quality due to time period effects. Statistics are broken down for VTE+ and VTE- (as an eligible event for the elderly surgery measure), and for with and without SES variables, in order to assess their effects on model quality with respect to discrimination and calibration.

In general: (a) model quality remains consistent when the 2011-2014 equations are applied to a unique dataset (2010) and when applied to subsets of data with an approximate 25% overlap; and (b) model quality is essentially

unaffected by the presence versus absence of SES in the prediction equation, while removal of VTE from the outcome appears to slightly decrease the fit with an increased discrimination. Because of the very large sample sizes studied here, a statistically significant Hosmer-Lemeshow statistic is not considered informative with respect to calibration.

Model	Model c statistic	2010 c statistic	Model HL	Model p-value	2010 HL	2010 p-value	Model Brier	2010 Brier
Elderly measure With VTE, Without SES	0.7586	0.7656	30.4886	0.0002	126.5938	0.0000	0.0888	0.1008
Elderly measure With VTE, With SES	0.7588	0.7659	31.1788	0.0001	131.9001	0.0000	0.0888	0.1008
Elderly measure Without VTE, Without SES	0.7714	0.7875	68.3632	0.0000	32.8440	0.0001	0.0771	0.0825
Elderly measure Without VTE, With SES	0.7715	0.7877	66.0084	0.0000	36.3637	0.0000	0.0771	0.0825

Jul 1, 2014 - Jun 30, 2015	C Statistic	HL	p_value	Brier
Elderly measure with VTE, without SES	0.7579	18.6091	0.0171	0.0872
Elderly measure with VTE, with SES	0.7580	16.9361	0.0308	0.0872
Elderly measure without VTE, without SES	0.7637	158.9828	0.0000	0.0819
Elderly measure without VTE, with SES	0.7638	162.3122	0.0000	0.0819

Jul 1, 2013 - Jun 30, 2014	C Statistic	HL	p_value	Brier
Elderly measure with VTE, without SES	0.7559	15.9363	0.0433	0.0873
Elderly measure with VTE, with SES	0.7561	16.6497	0.0340	0.0872
Elderly measure without VTE, without SES	0.7631	176.8872	0.0000	0.0820
Elderly measure without VTE, with SES	0.7633	183.0287	0.0000	0.0820

Jul 1, 2012 - Jun 30, 2013	C Statistic	HL	p_value	Brier
Elderly measure with VTE, without SES	0.7602	24.2472	0.0021	0.0877
Elderly measure with VTE, with SES	0.7605	25.3237	0.0014	0.0877
Elderly measure without VTE, without SES	0.7865	430.8442	0.0000	0.0671
Elderly measure without VTE, with SES	0.7867	429.6263	0.0000	0.0671

Jul 1, 2011 - Jun 30, 2012	C Statistic	HL	p_value	Brier
Elderly measure with VTE, without SES	0.7599	37.7755	0.0000	0.0951
Elderly measure with VTE, with SES	0.7601	37.6694	0.0000	0.0951

Jul 1, 2011 - Jun 30, 2012	C Statistic	HL	p_value	Brier
Elderly measure without VTE, without SES	0.7855	195.9585	0.0000	0.0740
Elderly measure without VTE, with SES	0.7857	188.3707	0.0000	0.0740

[Response Ends]

2b.27. Provide risk model discrimination statistics.

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

[Response Begins]

See 2b.26

[Response Ends]

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

[Response Begins]

See 2b.26

[Response Ends]

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

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

[Response Begins]

Unavailable

[Response Ends]

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

[Response Begins]

N/A

[Response Ends]

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

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

[Response Begins]

2022 Maintenance

The 3-predictor model provides effective control for differences in patient and procedure mix across hospital and yields fair hospital benchmarking.

[Response Ends]

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

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

[Response Begins]

none

[Response Ends]

3. Feasibility

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

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

[Response Begins]

Generated or collected by and used by healthcare personnel during the provision of care (e.g., blood pressure, lab value, diagnosis, depression score)

Coded by someone other than person obtaining original information (e.g., DRG, ICD-10 codes on claims)

Abstracted from a record by someone other than person obtaining original information (e.g., chart abstraction for quality measure or registry)

[Response Ends]

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

In other words, indicate whether data elements that are needed to compute the performance measure score are in defined, computer-readable fields.

[Response Begins]

Some data elements are in defined fields in electronic sources

[Response Ends]

3.03. If ALL the data elements needed to compute the performance measure score are not from electronic sources, specify a credible, near-term path to electronic capture, OR provide a rationale for using data elements not from electronic sources.

[Response Begins]

2022 Maintenance

There is an ongoing emphasis on collecting as many variables as possible from an EMR. However, it is unlikely that complete automation will be accomplished. This is the result of both technical limitations in each hospital's ability to automate EMR data downloads and because EMR variables might not have the same clinical definitions required by NSQIP.

[Response Ends]

3.04. Describe any efforts to develop an eCQM.

[Response Begins]

none

[Response Ends]

3.06. Describe difficulties (as a result of testing and/or operational use of the measure) regarding data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, time and cost of data collection, other feasibility/implementation issues.

[Response Begins]

- Data collection is a routine operation of SCRs (surgical clinical reviewers). Because of Covid's effects on hospital resources and efficiency of data collection, case locking (90 days after the date of operation) has been suspended on several occasions. Other than this, there have been no emergent issues in data collection.
- Except for the extension of data collection deadlines there have been no or feasibility issues affecting data collection and sampling.
- There have been no problems in our system's ability to maintain patient confidentiality with regards to both governmental regulations and contractual obligations.
- NSQIP costs are a concern for some hospitals. However, in more than 15 years of the programs existence ACS NSQIP has never implemented a price increase. While costs might seem excessive solely for "the" measure, it needs to be understood that the costs cover the entire benchmarking program and now allows them to participate in the Quality Verification Program free of charge .

[Response Ends]

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

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

Attach the fee schedule here, if applicable.

[Response Begins]

The cost for participation in Adult NSQIP is $26,000 annually; the cost to sites that qualify for the Small and Rural option is $10,000. Sites (except small and rural) must also support at least 1 full time, certified NSQIP data abstractor.

[Response Ends]

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.

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

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

4a.01. Check all current uses. For each current use checked, please provide:

Name of program and sponsor

URL

Purpose

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

Level of measurement and setting

[Response Begins]

Quality Improvement with Benchmarking (external benchmarking to multiple organizations)

[Quality Improvement with Benchmarking (external benchmarking to multiple organizations) Please Explain]

2022 Maintenance

Hospital-specific benchmarking results are sometimes shared with collaboratives that a hospital might be a member of. These collaboratives are either "blind", where each member's performance is reported anonymously among the group or "open", where each member's performance is identified to either collaborative leadership or to every member of the collaborative, depending on their agreement.

Quality Improvement (Internal to the specific organization)

[Quality Improvement (Internal to the specific organization) Please Explain]

2022 Maintenance

Results for the quality metric are made available to the hospital in quarterly reports and in an On-demand application.

[Response Ends]

4a.02. Check all planned uses.

[Response Begins]

Professional Certification or Recognition Program

Measure Currently in Use

[Response Ends]

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

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

[Response Begins]

NA

[Response Ends]

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

A credible plan includes the specific program, purpose, intended audience, and timeline for implementing the measure within the specified timeframes. A plan for accountability applications addresses mechanisms for data aggregation and reporting.

[Response Begins]

NA

[Response Ends]

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

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

[Response Begins]

2022 Maintenance

Measure results are included in NSQIP reports that are released quarterly in January, April, July, and October. In addition, participants can obtain more timely results (SAR reports are based on 1-year of data, collected 6-18 months before report release) using an "On-demand" application that uses a different methodology which relies on an historical equation which is updated annually and based on 4-5 years of data. Case data are accessed for these results as the record is marked as complete.

Cohen, M. E., Liu, Y., Huffman, K. M., Ko, C. Y. Hall, B. L. On-demand reporting of risk-adjusted and smoothed rates for quality improvement in ACS NSQIP. *Annals of Surgery,* 2016, **264**, 966-972.

[Response Ends]

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

[Response Begins]

2022 Maintenance

See 4a.06. Each report comes with complete documentation.

[Response Ends]

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

[Response Begins]

2022 Maintenance

There is no mechanism in place that provides feedback specific to the NQF measure vice feedback with regard to participation in the entirety NSQIP including the measure. Feedback channels are

[Response Ends]

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

[Response Begins]

2022 Maintenance

NSQIP participants provide comments and feedback with regard to sampling requirements and clinical definitions. NSQIP participants (data collectors, nurses, and surgeons), either independently or as members of advisory groups, actively participate in advancing the clinical validity of our measures.

[Response Ends]

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

[Response Begins]

2022 Maintenance

We rarely get feedback from non-participants.

[Response Ends]

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

[Response Begins]

2022 maintenance

As described previously in this submission, definitions of predictors have evolved for this measure. These changes are influenced by participant feedback.

[Response Ends]

4b.01. You may refer to data provided in Importance to Measure and Report: Gap in Care/Disparities, but do not repeat here. Discuss any progress on improvement (trends in performance results, number and percentage of people receiving high-quality healthcare; Geographic area and number and percentage of accountable entities and patients included). If no improvement was demonstrated, provide an explanation. If not in use for performance improvement at the time of initial endorsement, provide a credible rationale that describes how

the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.

[Response Begins]

2022 Maintenance

We have observed continuous reduction in NSQIP-tracked outcomes since program inception. However, the use of "the" measure and participation in the entirety of NSQIP are confounded. Improvement in outcomes cannot be assigned independently to the measure.

[Response Ends]

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

[Response Begins]

2017 Maintenance

Based upon experience with ACS NSQIP data collection, there are very few problems with errors or inaccuracies. Data collectors in the ACS NSQIP receive extensive training and support for accurate data collection. In addition, data collectors are audited for inter-rater reliability and are held to a 95% or better concordance rate for all variables. Additionally, chart audits have been planned in accordance with CMS stipulations for measure participants who are not ACS NSQIP participants.

[Response Ends]

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

[Response Begins]

None noted

[Response Ends]

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.

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

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

(Can search and select measures.)

[Response Begins]

[Response Ends]

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

(Can search and select measures.)

[Response Begins]

[Response Ends]

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

[Response Begins]

N/A

[Response Ends]

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

[Response Begins]

Yes

[Response Ends]

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

[Response Begins]

N/A

[Response Ends]

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

Provide analyses when possible.

[Response Begins]

NA - different target population

[Response Ends]

Appendix

Supplemental materials may be provided in an appendix.:

No appendix

Contact Information

Measure Steward (Intellectual Property Owner): American College of Surgeons

Measure Steward Point of Contact: Ali, Sameera, sali@facs.org

Measure Developer if different from Measure Steward: American College of Surgeons

Measure Developer Point(s) of Contact: Cohen, Mark, mcohen@facs.org

Cohen, Mark, mcohen@facs.org

Ali, Sameera, sali@facs.org

Additional Information

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

[Response Begins]

No appendix

[Response Ends]

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

Describe the members' role in measure development.

[Response Begins]

Clifford Ko

Sameera Ali

Bruce Hall

Mark Cohen

Yaoming Liu

This group used ACS NSQIP data to develop the statistical risk-adjusted model on which this measure is based. The workgroup also reviewed and summarized the literature that supports the importance of using this measure to as a tool to improve surgical quality.

[Response Ends]

3. Indicate the year the measure was first released.

[Response Begins]

2011

[Response Ends]

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

[Response Begins]

January 2022

[Response Ends]

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

[Response Begins]

Annual

[Response Ends]

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

[Response Begins]

January 2023

[Response Ends]

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

[Response Begins]

N/A

[Response Ends]

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

[Response Begins]

N/A

[Response Ends]

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

[Response Begins]

N/A

[Response Ends]